Swipe and Tell: Using Implicit Feedback to Predict User Engagement on Tablets

KLAAS NELISSEN, KU Leuven
MONIQUE SNOECK, KU Leuven
SEPPE VANDEN BROUCKE, KU Leuven
BART BAESENS, KU Leuven and University of Southampton

When content consumers explicitly judge content positively, we consider them to be engaged. Unfortunately, explicit user evaluations are difficult to collect, as they require user effort. Therefore, we propose to use device interactions as implicit feedback to detect engagement.

We assess the usefulness of swipe interactions on tablets for predicting engagement, and make the comparison with using traditional features based on time spent.

We gathered two unique datasets of more than 250,000 swipes, 100,000 unique article visits, and over 35,000 explicitly judged news articles, by modifying two commonly used tablet apps of two newspapers. We tracked all device interactions of 407 experiment participants during one month of habitual news reading.

We employed a behavioral metric as a proxy for engagement, because our analysis needed to be scalable to many users, and scanning behavior required us to allow users to indicate engagement quickly.

We point out the importance of taking into account content ordering, report the most predictive features, zoom in on briefly read content and on the most frequently read articles.

Our findings demonstrate that fine-grained tablet interactions are useful indicators of engagement for newsreaders on tablets. The best features successfully combine both time-based aspects and swipe interactions.

CCS Concepts: • **Computing methodologies** → *Learning from implicit feedback*; • **Applied computing** → *Publishing*; • **Human-centered computing** → *Tablet computers*;

Additional Key Words and Phrases: User Engagement; Implicit Feedback; Tablets; Dwell Time; Touch Interactions; Newspaper; Online News, Content Ordering; Briefly Read Content; Frequently Read Content.

ACM Reference format:
Klaas Nelissen, Monique Snoeck, Seppe vanden Broucke, and Bart Baesens. 2018. Swipe and Tell: Using Implicit Feedback to Predict User Engagement on Tablets. *ACM Transactions on Information Systems* 1, 1, Article 1 (February 2018), 37 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

User engagement is defined as the quality of the user experience that emphasizes the positive aspects of interacting with an online application [Lalmas et al. 2014]. Users are engaged when they appreciate the content to which they have given their attention. Identifying when users are engaged is interesting because it provides content creators with insights on how their products are used, and so it can be used to improve the offering towards users. At a small scale, we could just ask users to judge the content they consume and thus get accurate explicit user evaluations. And although explicit user judgments are the best measures for assessing engagement, it requires a high cognitive effort [O'Brien and Toms 2010]. Moreover, in online applications, this explicit feedback

The authors would like to thank Twipe for their cooperation with this research. This work was facilitated by iMinds and funded by VLAIO (grant number 140655).
Correspondence concerning this article should be addressed to Klaas Nelissen (e-mail: Klaas.Nelissen@kuleuven.be).
2018. 1046-8188/2018/2-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

is always given voluntarily, and thus not scalable to a large numbers of users. This study fits in the research looking for better proxies for explicit user evaluations which can be used to improve large-scale measurements of user engagement.

One way user engagement has been measured at large scale is by tracking how much time users spend with content. But the time spent (i.e., dwell time) does not necessarily indicate appreciation. A user may spend 30 seconds reading the first half of a text attentively, or may be skimming through the whole text, scanning for relevant information. So there is a need for finer measures for user engagement, and this has been proven successful in web search on computers where mouse interactions and scrolling behavior could be used for identifying document relevance [Guo and Agichtein 2012]. User engagement differs from relevance because engagement emphasizes also the hedonic and affective aspects of the user experience, while relevance focuses on need satisfaction or goal accomplishment. Conceptually, engagement and relevance are closely related, which is why we can use insights from relevance research to inform our study of engagement. We discuss these terms and theid differences more in-depth in the literature review. A better web search result ranking could be achieved on mobile devices by taking into account fine-grained swipe interactions [Guo et al. 2013b; Huang et al. 2011]. Recent research has shown that users' experiences are different on different devices, and earlier gained insights might not be transferable across devices [Huang and Diriye 2012]. While other studies have primarily focused on web search on computers, this study extends the current research to the context of news reading on tablets.

In web search, the order of presenting the results to a query has a very large impact on the click through rate [Agichtein et al. 2006b]. In a newspaper, content is also presented in a ranked order chosen by the editors. This decision about when to present which content to the reader is a key aspect of the newspaper creation process, which the editors spend a lot of time and effort on. It is therefore interesting to look into the usefulness of features based on the content ordering and the structure of the newspaper for predicting engagement, in the context of a tablet app for a digital newspaper.

Another interesting question to ask is how to detect engagement for content which is read for only a short period of time. In general, when considering interactions or experiences of a short duration, the approach of using *time spent* will not work anymore and alternative approaches are required. For briefly read content, features based on time spent will have a reduced range and variance, so they will contain less information. Different interaction features might play a different role in this use case. Also, for each piece of content a reader comes across, the reader makes an (unconscious) decision about whether to spend more time with it or not. Editors are especially interested in those situations where a reader only interacts briefly with some content, but still judges that content positively. Because editors optimize for engaging content, it is interesting to investigate which interaction behaviors lead to readers judging content positively which they have only read briefly.

The final question we study concerns the difference between articles which are frequently read and those which are not. We repeat the analysis for the 25% most frequently read articles. From discussions with the newspaper editors, we learned that they spend relatively more time analyzing and discussing these more popular articles, trying to find out why these articles work so well. The most frequently read articles also function as a common divisor across the whole user population, thereby giving editors insight into the preferences of their reader base. The most important features for predicting engagement with these most frequently read articles might also be different.

The research questions of this paper are:

R.Q. 1: How do fine-grained swipe interactions (as implicit feedback features) compare to time-based features in terms of performance for predicting user engagement in the context of news reading on tablets, and which are the most important features?
R.Q. 2: How do features related to the structure of the newspaper and the ordering of the content perform for predicting engagement?
R.Q. 3: How useful are fine-grained swipe interactions for predicting engagement with briefly read content?
R.Q. 4: To what extent does the performance of the models for predicting engagement change when we consider only the most frequently read articles?

To find an answer to these research questions, we did experiments with people who read a digital newspaper on a tablet app. We instrumented two apps to track every user interaction, added an in-app feedback mechanism, and asked users to give feedback when they found certain content engaging. For each article in the newspaper, users could give a thumbs up or down, so we obtained a large set of explicitly judged articles.

We consider a user to be engaged with an article when she gives a thumbs up on that article. Admittedly, this is a simplistic behavioral measure which functions as a proxy for user engagement and which does not capture the holistic nature of user engagement as discussed by O'Brien and Toms [2008; 2010]. However, this metric does satisfy our requirements of allowing large-scale measurements of user engagement which are scalable to all users, and it demands almost no user effort, so the metric allows users to quickly give a thumbs up to newspaper articles they were just scanning over. Even more important, a simple behavioral metric disrupts the regular reading experience as little as possible. We further address in the methodology section why choosing for a simple operationalization of engagement is the best option for this study and why alternative methods based on a lengthy survey are not feasible.

We identified a large number of interaction features to capture the user behavior while reading, and used these features in logistic regression models to predict whether a user will judge an article positively or not.

In summary, we make the following contributions:

- We extend the current research on scalable measurements for user engagement to the context of news reading on tablets.
- We contrast the usefulness of device interactions as implicit features versus time-based features for predicting user engagement.
- We illustrate that features based on the order in which content is presented and the general newspaper structure are useful for predicting user engagement.
- We discuss user engagement predictions for briefly read content and for the most frequently read articles, showing that different types of features perform differently in each of these two specific settings (which have not been analyzed separately before).

2 RELATED WORK

Song et al. [2013a] make the point that user behavior on tablets is not only different from user behavior on computers, but also from user behavior on smartphones. They suggest that each device should be treated differently, and that insights are not necessarily transferable across devices. Content which causes engagement on computers or smartphones is not necessarily also engaging on tablets [Lu et al. 2014]. Huang and Diriye [2012] argue in a position paper that touch events have a different meaning than cursor events but that they have great potential in helping to better understand user experiences. They propose to focus on tracking the viewport. This is the part of

the page the user is currently seeing, and is more useful on smaller screens such as smartphones and tablets. Several features included in our analysis are based on the viewport.

2.1 The usefulness of device interactions

There is little empirical research which specifically focuses on touch interactions for detecting user engagement, or more generally, for identifying positive aspects of the user experience. Most closely related to our work is the study by Guo et al. [2013a], which shows that web search result rankings can be significantly improved by taking into account touch interactions. The authors conducted an experiment where users were asked to answer a number of questions by searching the web, and while every touch interaction during the search was captured, the users also explicitly rated every page they visited. Two of the most useful features in their study are the swipe frequency (which is the number of swipes on a page divided by the dwell time on that page) and the maximum inactive time between two touch interactions. They find that more and faster swipes are negatively correlated with document relevance, as they indicate scanning behavior. In contrast, slow swiping and long periods of inactive time suggest that users are paying attention and actively reading the current web page.

There are more studies which do not specifically focus on touch interactions, but show the value of implicit interactions for estimating appreciation of content, document relevance, or user engagement. Several studies in both the domains of information retrieval [Agichtein et al. 2006a; Fox et al. 2005; Guo and Agichtein 2012; White et al. 2005] and recommender systems [Konstan et al. 1997; Lee and Park 2007; Liu et al. 2010] have shown that implicit interactions are useful for distinguishing document relevance. Based on implicit feedback, Guo and Agichtein [2008] could in one of their earlier studies identify whether a searcher had an intent to purchase or was just browsing for information. In another study by the same authors, they prove that incorporating post-click searcher behavior (such as scrolling and cursor movements) in addition to dwell time and clickthrough statistics can improve estimates of document relevance [Guo and Agichtein 2012]. Their analysis asserts that slow gestures might be indicative of reading, while faster mouse gestures might characterise a navigational pattern to locate certain information of interest in the text. Agichtein et al. [2006b] show that implicit feedback can be of even more value if the features are modeled as deviations from expected user behavior. We also include deviational features in our current study.

Other studies show that using fine-grained mouse interactions offer a scalable way to infer user attention on web pages [Claypool et al. 2001; Huang et al. 2011]. Huang et al. [2011] did a study where they correlate cursor movements on web pages with explicit relevance judgments of users. They show that incorporating these fine-grained cursor interactions can improve estimates of document relevance. In their experiments, the mouse hover rate is the feature which correlates best with human relevance judgments. In contrast, the duration of mouse hovers correlates negatively with relevance in their study, while in other studies such as the one by Claypool et al. [2001], cursor travel time is a positive indicator of web page relevance. Unfortunately, features relating to 'hovering' do not have their equivalent in terms of tablet interactions.

Navalpakkam and Churchill [2012] use mouse cursor interactions to predict whether the reading experience of the user is pleasant or not significantly better than normal. They report that long and frequent mouse visits on text are strong predictors of an unpleasant experience. Speicher and Gaedke [2013] do a similar study which results in their end-to-end system TellMyRelevance. The system learns relevance models by automatically tracking and analyzing client cursor interactions.

Arapakis et al. [2014a] model a large set of features based on mouse interactions with the goal of developing a taxonomy of mouse patterns for determining interestingness of web pages. They

include more than 60 features describing how the mouse was used. Only features based on speed, and minimum, average, and total distance are significant. The already mentioned study by Guo and Agichtein [2012] finds similar results, where frequency and speed correlate with document relevance. Lagun et al. [2014] took this a step further, using dynamic time warping to automatically identify cursor motifs (frequent subsequences) which could then be used as features for more accurate estimations of relevance. Recently, Liu et al. [2015] improved on this study by coming up with distance- and distribution-based strategies for extracting the motifs, which predict satisfaction with search results even better, even when using a smaller number of motifs. This holds true even when predicting satisfaction for previously unseen users or search queries, which is important for practical applications. Shapira et al. [2006] find that mouse travel distance is a worse indicator than the ratio of mouse movement to reading time for document relevance.

The evidence of using only page dwell time for inferring relevance shows mixed conclusions [Fox et al. 2005; Guo et al. 2013a; Lagun and Lalmas 2016; Yi et al. 2014]. The correlation between time spent and relevance is often, but not always, significantly positive [Liu et al. 2016]. Early research shows that there is a strong tendency that users spend more time on interesting rather than uninteresting news articles [Claypool et al. 2001; Morita and Shinoda 1994]. However, dwell time is for example not the best indicator for page quality in the study done by Shapira et al. [2006]. Sundar, Bellur, Oh et al. [2014] suggest that the relationship between higher levels of interactivity and favorable attitudes and behavioral intentions could indeed be curvilinear. They show show through experiments that different on-screen interaction techniques (such as clicking, zooming, sliding, and so on) differ in their ability to influence the occurrence of engagement, as measured by cognitive absorption. The same study finds that the concept of user engagement cannot be reduced to a measurement of the volume of user actions. That is, a higher amount of user action may according to this study not always be indicative of higher engagement. In another study by Oh and Sundar [2015], they show that structural interactivity elements (modality and message interactivity) of a website have a positive effect on the occurrence of engagement. Sundar et al. [2016] recently expanded on this study by framing message interactivity, operationalized as user interaction history, in a dialogue and conversation context between user and system. They found that perceptions of contingency are critical for user engagement with a site, and that engagement in turn predicts user attitudes and behavioral intentions towards that site. This study again suggests that it is not the amount of interactions which creates enhanced engagement, but rather the perceptions of contingency [Sundar et al. 2016]. This research suggests that interaction tools translate into greater user engagement when users find the interaction to be simple, natural, and/or intuitive.

In summary, past research suggests that using fine-grained interactions in addition to features based on *time spent* proves to be useful for explaining document relevance on computers. However, none of these studies which use implicit feedback make the difference between briefly or long read content, or focus on the most frequently accessed items. Most previous experiments took place on computers.

2.2 Defining and measuring user engagement

O'Brien and Toms [2008; 2010] did the fundamental work of constructing a good definition for user engagement as well as developing a valid and reliable 31-item survey. They define user engagement as the quality of the user experience, characterized by user engagement attributes. They identified six distinct attributes of engagement: perceived usability, aesthetics, focused attention, felt involvement, novelty, and endurability. Their findings indicate that these attributes are highly intertwined, and that engagement is both a process and a product of interaction which can vary in intensity over the course of an experience. O'Brien also situates these findings in the context of mobile devices in

a different study [O'Brien et al. 2013]. Sundar et al. [2014] state that engagement is a multifaceted concept which is often equated with absorption, immersion, transportation, flow, and presence. Oh and Sundar [2015] consider both cognitive absorption and elaboration (number of thoughts on the current experience) as indicators of engagement. Engagement is conceptually a holistic and experience-based framework which exceeds usability, it characterizes to what extent an application can provide a pleasurable or memorable experience [O'Brien 2011].

Other research suggests that there is not one best approach to measure user engagement, but that the most suitable measurement method depends on the online experience which is being studied [Lehmann et al. 2012]. The overview by Lalmas, O'Brien, and Yom-Tov [2014] describes three different measurement methods, each with its own advantages and drawbacks: self-reports, physiological signals, and behavioral metrics.

Surveys are subjective and hard to administer at massive scale. Physiological signals such as EEG or eye-trackers offer the most objective measurement method, but the need for specialized equipment limits their practical use outside research [Lalmas et al. 2014]. Only behavioral metrics allow researchers to collect data from all users of a service with almost no user effort, which is one of the requirements for our current study. These behavioral metrics are unable to explain *why* users find something engaging, they can only act as a proxy for user engagement [Lehmann et al. 2012].

Using behavioral metrics as proxies for user engagement is also done by Song et al. [2013b] and Drutsa and Serdyukov [2015]. In Song et al. [2013b], the authors develop a machine learning model which can predict drops in user engagement (as measured by behavioral metrics) on the long term by having previously purposefully degraded the relevance of returned web search results. The starting point of another study by Lagun and Lalmas [2016] is the acknowledgement of the limitations of dwell time as a metric for user engagement, specifically because dwell time can not tell whether a user is paying attention or not. Using viewport data from a computer they come up with four scalable behavioral metrics which capture different levels of intensity of engagement: bounce, shallow engagement, deep engagement and complete engagement. Their unit of analysis is one news article, but there is no ground truth of engagement provided by a user. Another recently proposed behavioral metric by Dupret and Lalmas [2013] is absence time, which is defined as the time between two user visits. While the results of this study are promising, this metric is not relevant for our current research because we do not consider engagement levels over different reading sessions.

Arapakis et al. [2014b] investigate user engagement in online news on computers. They do not use any behavioral metrics based on user interactions, but instead use eye tracking as the objective measure for user engagement, and use surveys to determine the interestingness of news articles, among other things. They find that the level of focused attention is determined by the perceived interestingness of the news article. This finding is corroborated in a study by McCay-Peet et al. [2012], where a user's self-reported level of interest in a topic is found to be a good predictor for self-reported focused attention. Wu, Liu, Su, et al. [2017] have shown that it is feasible to use physiological signals in the form of features derived from electrodermal activity to predict search satisfaction and explain mobile shopping behaviors.

Closely related to engagement is relevance, which can be defined as the perceived amount of useful information a user acquired from a document, or the ability to retrieve material that satisfies the needs of the user. Saracevic [2007] notes in his review on relevance that a general relevance theory might be too complex and complicated to sort out theoretically all at once, and that relevance is not one thing but many and depends on interpretation. He proposes to concentrate on a limited number of key factors or manifestations of relevance: system or algorithmic relevance, topical or

subject relevance, cognitive relevance or pertinence, situational relevance or utility, and affective relevance [Saracevic 2007].

Another closely related concept is document usefulness, which was proposed by Belkin, Cole, and Liu [2009]. Usefulness is grounded in the nature of information seeking, which is said to take place in the circumstance of having some goals to achieve or task to complete [Belkin et al. 2009]. Cole, Liu, Belkin et al. [2009] then conceptualized usefulness as explicitly considering one session as a whole, thus considering usefulness to be more general than relevance. They explain that measuring usefulness depends on the specification of a task or goal, of which the achievement can be measured [Cole et al. 2009]. Document usefulness proved to be a good measure for optimizing information retrieval performance in [Liu et al. 2012]. Nevertheless, Liu and Belkin [2015] demonstrate that inferring the usefulness of a document should be tailored toward individual tasks and users. Jiang et al. [Jiang et al. 2017a] show through doing a lab study that usefulness as a metric better correlates with six user experience measures rather than using just topical relevance. The six user experience measures are satisfaction, goal success, frustration, task difficulty, system helpfulness, and total effort spent. Their results suggest that it is almost always helpful to complement either relevance or usefulness with one of the four alternative aspects of an experience (novelty, understandability, reliability, and effort) to better correlate with user experience measures [Jiang et al. 2017a]. The holistic approach of user engagement attempts to encompass all these concepts. User engagement is a measure which tries to embrace these different aspects of the experience in one concept. Conceptually, the findings from relevance and usefulness research can inform our study on engagement because the concepts are closely related.

Jiang, He, Kelly, and Allan [2017b] recently proposed ephemeral state of relevance (ESR) as a metric. It is defined as the amount of useful information a user acquired as assessed just after examining the result under a natural condition at a particular moment of a search process. It is assessed by real searchers in real time, by the criterion of the document being useful to the problem at hand. It captures the real-time state of mind and perceptions of a user. They consider ESR as a particular implementation of Belkin et al.'s [2009] evaluation model where the second level measures the usefulness regarding 'each interaction'. In the context of ESR, relevance and usefulness are interchangeable [Jiang et al. 2017b].

In the user experience of reading online news, previous research has shown that users have a high desire for choice and novelty of content, and that the content itself is an important quality of users' engagement when interacting with news [O'Brien 2011]. Novelty and personal interest are two specific attributes of user engagement which are more important in the context of online news [O'Brien 2011]. In a later study by O'Brien and Lebow [2013], the authors argue that past research has shown that news consumption is not only driven by the desire to locate specific information, but also to stay up-to-date on current events, or to engage in leisure activities and also out of habit.

News reading is not necessarily related to a task. A particular reading experience does not have a specific or amorphous goal but can be started with the goal of having fun, passing the time, and so on. Although users might want to locate specific facts or enhance their understanding of a specific topic, this is not a requirement for having an engaging experience.

O'Brien and Lebow [2013] propose a broader conceptualization of experience, viewing information encounters in a more experiential manner, in which pragmatic and hedonic aspects blend together. They say that if we want to understand information interaction as an experience, that means that we must look at more than outcomes of 'success' (that is, did the user locate the information to satisfy the information need?) [O'Brien and Lebow 2013]. Finally, they advocate to use metrics which capture both the pragmatic (e.g. usability) and the hedonic (e.g. fun, engagement,

absorption) aspects of information interactions in online news interactions [O'Brien and Lebow 2013].

News reading is partially a hedonistic experience, while information retrieval metrics such as relevance and usefulness focus on success of a task. Taking a holistic approach allows different users to underscore different attributes of user engagement, as they might have different preferences.

In summary, simple behavioral metrics are frequently employed as proxies for user engagement. In fact, when the measurement method for user engagement is required to be scalable to all users, behavioral metrics are the only viable method. User engagement is harder to measure, as it covers several distinct aspects of the user experience, is formed in the long run, and often does not follow from a goal-oriented experience, which also makes it harder to evaluate [Lalmas et al. 2014].

The most advanced studies in measuring user engagement try to combine different measurement methods to better measure engagement. O'Brien and Lebow [2013] were among the first to set up a study which employed this mixed-methods approach by including both surveys, behavioral metrics, and physiological signals. Mathur, Lane and Kawsar [2016] also combine EEG signals, self-reported perceived engagement scores, and eventually also contextual features automatically derived from smartphones to successfully develop a machine learning model which can detect different levels of engagement.

Our work builds on previous research connecting explicitly expressed user engagement with device interaction behavior. To the best of our knowledge, it is the first to consider mining touch interaction data on tablets in the context of news reading, to take into account the ordering of the content, and to investigate engagement on briefly read articles and on frequently read articles.

3 METHODOLOGY

As an operationalization of user engagement, we use the presence of a user's explicit feedback on an article as the positive outcome of a binary feature. By giving a thumbs up, the user indicates appreciation. One observation in our dataset is one article visit by one user. How we obtained the explicit judgments in the app is further explained in the section on the experimental set-up. The binary dependent feature says whether the users judged the article positively, or not. Of course, this is a coarse and short term operationalization, which can only function as a proxy for user engagement.

The focus of the current paper is on investigating which type of variables are the most useful to predict the occurrence of engagement in a practically applicable industry setting. Doing this does not require to further characterize the engagement or speculate about the reasons why the user had been engaged. So while we recognize that we use a simple, binary evaluation of engagement and in reality, engagement with a news article is not evaluated by users as being binary but can differ in intensity, our chosen operationalization does seem appropriate to achieve the objectives set out in this study. If we would want to explore a theory about when engagement would occur for example, a more complex measurement of engagement would be necessary.

As the current paper is part of a larger research project executed in collaboration with publishing companies, it was imperative that the experiments were designed so that the results would have as much practical applicability as possible. The methodology needed to be able to be implemented in practice and to be scalable to many users.

To maximize the practical usefulness of our findings for industry practitioners, we opted for a longitudinal naturalistic study. By doing so, we aimed to realize a large external validity of our conclusions.

In an ideal experiment, different variations on the in-app explicit feedback collection method could be tested with different large enough user groups, and the effect of the method on the results

could be estimated by including it as a variable in the analysis. This would enable a multi-method analysis of the results, as recommended by [Podsakoff et al. 2003]. However, this was not feasible for us as doing two quantitative experiments and modifying two apps took already a lot of resources and time to develop. So although we acknowledge there might be some method bias present in our study, we tried to keep it to a minimum by implementing a mechanism that resembles the natural setting of a reader as close as possible but still integrates explicit engagement evaluations.

A limitation of using users' explicit evaluations is that no ideal solution exists to collect this feedback without bothering users. In fact, it would be impossible to conduct the study if the method would require to not bother users at all. As is typical for behavioral research experiments, any analysis which relies on user-supplied feedback needs to implement a mechanism which will bother users to a certain extent. We would not have been able to observe the outcome measure without instrumenting the apps with some type of in-app feedback mechanism.

Given this limitation, we opted for a simple binary engagement measure, which allows us to stay as close as possible to the natural setting in which these reading experiences would normally take place in an uncontrolled environment, while still integrating explicit engagement evaluations. Such a behavioral measure is easily scalable to all users. We wanted our method for evaluating engagement to disrupt the regular reading experience as little as possible, while still allowing to collect explicit feedback. Edwards and Kelly [2016] note that although behavioral measures can be ambiguous, the physical manifestation of engagement is an important part of engagement, and it provides a useful and unobtrusive way to operationalize engagement. This means that the mechanism needed to require not much user effort or time. As Jiang et al. [2017a] note, collecting in situ judgments and user interaction behaviors together poses a challenge to experiment design. They note that collecting accurate in situ judgments often require multi-item measurements, but that interrupting participants for in situ judgments breaks the flow of search session and can affect subsequent interaction behavior. They propose to make compromises in experiment design by simply using one question to measure each user experience measure.

We opted for engagement as our measure rather than relevance, usefulness, or satisfaction because we believe the engagement concept best captures the holistic aspect of the digital news reading experience, which includes hedonic and non-goal directed elements, such as fun and novelty. In the context of digital news, users' motivations go further than completing an information task. User engagement allows to capture both utilitarian and hedonic aspects of the experience. Kelly [2009] also highlights the need to also capture affective responses when measuring user experiences. Because we did a longitudinal naturalistic study and consequentially limited the engagement evaluation to one question, it is possible that users take a mental shortcut and fuse different aspects of the user experience in their engagement evaluation.

The methods we employed were designed with the goal to change as little as possible to the regular reading experience, so that the method itself would not introduce variance in the outcome measure. Besides the explicit feedback collection mechanism, we did not deliberately introduce any other condition to introduce variance in the outcome measure. We did not further experimentally manipulate the regular reading experience. To ensure the unobtrusiveness of the method as much as possible, a preliminary version of the in-app feedback mechanism has been tested with a panel of readers, who were interviewed after a reading session to get qualitative feedback on how they experienced the in-app survey and to what extent the collected data would correctly reflect their state of mind if the mechanism was used in an uncontrolled environment outside a lab.

The simple engagement evaluation mechanism also satisfies the following requirements: it keeps the barrier to give feedback as low as possible by minimizing user effort, it allows users to quickly give feedback in a matter of seconds without disrupting the reading experience (so even while

scanning an article), it is scalable to many users, it is readily available to users and not hidden in a menu, and it is the same for every article so users quickly learn the mechanism and know what to expect. It also strikes the right balance between attracting enough attention so readers would not occasionally forget to give feedback, while it does not hinder reading the article too much or does not transform the normal reading experience to which the participants were already accustomed in a major way.

If the explicit feedback collection mechanism would have been more complicated because it had more questions and thus needed more time to complete, the experiments with a duration of one month would deviate more from how these reading experiences would normally take place in an uncontrolled environment. Filling in a survey with even a small number of questions would already interrupt the reading experience too much, and the experiments would have started to resemble more of a lab study, which would make our results less applicable in practice.

We tried to diminish the threat of method bias and made an informed choice for the explicit feedback collection methodology by letting participants give feedback on different possible implementations of the in-app explicit feedback collection mechanism during the pilot studies before each quantitative experiment. The pilot studies are further explained in the section on the experimental set-up.

One example of how we tried to minimize how the in-app feedback mechanism influenced the behavioral interactions variables is to have multiple locations to indicate engagement for the same article (the title, image, and at the end of the text of the article). Because of this, users do not have to scroll back to the top of the article just to indicate that they were engaged with the article that they just finished reading.

As editors of a newspaper are most interested in insights about user engagement on the level of one news article, we measured engagement with one article as a whole. Moreover, while we can calculate the values for the implicit interaction variables for one article, this would be much harder to implement if we wanted to calculate these variables' values for every title, image, and the text of an article separately. Several of these variables would not have a useful interpretation if they would be calculated for one title or one image, so we suspect they would not be of much use to predict engagement (e.g. the number of swipes on a title would not make much sense, as its value would always be zero). It would not be possible to investigate our research question about which type of variables is best for predicting engagement if we would do a separate analysis for each title, image, or text of an article.

Our study could suffer from social desirability bias when users do not indicate engagement with articles which they actually do find interesting in reality (this might be the case for gossip or human interest articles). This would be represented in our study by false negatives, of which the total number is quite small (as shown in the results).

We followed the suggestions of Podsakoff et al. [2003] to keep the items simple, clear, and specific, and to choose a short scale because it reduces some bias which could be produced by respondent fatigue or carelessness. Tourangeau et al. [2000] note that one of the most common problems in the response process is item ambiguity and we adopted their recommendations to avoid vague concepts, keep questions simple, specific, and concise, use focused questions, and avoid complicated syntax. As this was a longitudinal study of one month, and the study took place in participants' natural environment and not in a lab setting, no method effects could have been produced by the measurement context. We also do not use self-report measures as the sole type of data for our study, which is another potential major source of common method variance. We try to link explicit indications of engagement with implicit interaction with the tablets. We use self-reported engagement and actual behavior instead of self-reports alone.

Podsakoff et al. [2003] recommend to obtain measures of the dependent and independent variable from different sources as a technique to control method bias. Although our dependent and independent variables are collected during the same reading experience, engagement is self-reported while the independent variables are based on actual behavior. We ensured respondent anonymity and reduced evaluation apprehension by assuring respondents that there are no right or wrong answers and that they should indicate engagement as honest as possible.

Because we measure actual behavior, the variables collected by the device interactions are objective. The interactions a user has with the device are the same, irrespective of whether the in-app feedback mechanism is present or not. That is, the addition of the in-app feedback mechanism will not have a large effect on whether engagement would potentially occur or not. Intuitively, we also think that the different types of independent variables are not severely impacted by the presence of the in-app feedback mechanism. We believe that the presence of the in-app feedback mechanism has a rather small effect on influencing users' behavior to for example spend more or less time on an article, to swipe more or less on an article, read more articles, or take more time until interacting with the article.

The in-app feedback mechanism is kept subtle and unobtrusive, so that it would not cause users to find an article engaging which they would have not found engaging if the in-app feedback mechanism would not have been there, or vice versa, that the presence of the in-app feedback mechanism would cause users to not find certain articles engaging which they would have considered to be engaging without the in-app feedback mechanism.

Both the operationalization of engagement and the implementation of the in-app feedback mechanism are chosen such that the users' swipe behavior would differ as little as possible from how it would have been if there would not have been an in-app feedback mechanism present.

Although we make no assumptions about why engagement is occurring in the current paper, we believe it depends on factors such as the characteristics of the current article, its position in the newspaper, the interests and profile of the current reader, and so on. We believe that whether a user will in reality find an article engaging or not, will depend more on such factors than on whether there was an in-app feedback mechanism present or not. We believe that in our paper, the variance in the occurrence of engagement is not to a large extent influenced by the measurement method.

Malhotra, Kim, and Patil [2006] find that method bias in the IS domain is not as serious as that found in other disciplines, and that even when method bias is present, the biases are not substantial.

In an ideal setting, we would have had the resources to set up different groups of experiment participants which each used a different version of the modified app with a different implementation of the in-app feedback mechanism, and another group of which we only collected the interaction data but where we did not add the explicit in-app feedback mechanism. Such a set-up would have allowed us to estimate the effect of adding the method to collect explicit feedback, effectively constructing a valid baseline of the experience. We would then be able to control for the method to estimate the effect it has on the independent and dependent variables [Kelly 2009]. However, given the practical circumstances we did not have the manpower nor the financial resources to set up an experiment three times the size of ours (which already lasted twice one month with in total 407 participants) or to develop multiple versions of the in-app feedback mechanism. One could argue that the in-app feedback method has biased our results to a certain extent. However, in the hypothetical situation that we would have been able to set up a control group of 400 users tracked for a period of one month, we would likely have had a difficult time to distinguish variance in the occurrence of engagement caused by the in-app feedback mechanism from variance in the occurrence of engagement caused by the reading experience and the news articles themselves.

Table 1. Time-based features and strictly implicit features used for modeling.

Feature Name	Description
Time-based features	
timeOnArticle	The time in seconds a user spent on the article.
timeOnPage	The total time in seconds spent on the current page where the article is situated.
timeSpentNextPage	The total time in seconds a user spent on the next page.
timeSpentPrevPage	The total time in seconds a user spent on the previous page.
isNextPageRead	Whether the next page is read, taking into account the number of words on that next page.
isPrevPageRead	Whether the next page is read, taking into account the number of words on that previous page.
Strictly implicit features	
articleCompleteness	A % giving the proportion of an article the user has *seen* by scrolling down vertically.
weekend	Whether the session took place during the weekend or not.
nrSwipesArticle	The number of swipes on an article.
nrSwipesPage	The total number of swipes on the current page where the article is situated.
timeToFirstInterPage	The time in seconds it took until the user first interacted with the current page. We consider this to be an implicit feature because the time-based features do not take into account any user interactions.
timesViewnThisPage	The number of times the user visited this page.
tappedTeaser	Whether the user tapped on a teaser to jump to this article, or not.
nrSessionsNewspaper	The total number of distinct reading sessions on this newspaper.
sessTimeOfDay	Categorical feature saying when the session was taking place; possible values: morning (until 10AM), day (until 5PM), evening.
daysSincePrevSess	The number of days since the user's previous session.
isImageOpened	Whether the user tapped on an image in this article, or not.

We use logistic regression models to predict whether a user was going to be engaged with an article. We also tried random forests, but this method did not improve the results. We chose logistic regression because it is fast, easy to integrate in internal company tools, and the coefficients of the model offer an intuitive interpretation for feature importance. This makes it easier to communicate the results of the models to a non-technical audience such as editors and journalists. As the dataset is large enough, we could evaluate the predictive performance of the model by doing out-of-time-validation, which is the strongest way to test predictive models [Baesens et al. 2015]. We keep the last 25% of the data separate for testing, which is the fourth week of the month-long study. As the newspapers we studied get released every day of the week except on Sunday, the test set includes only articles which were not seen by any user before. Although in some of the models there is a clear class imbalance, using the SMOTE resampling technique [Chawla et al. 2002] did not significantly improve the results.

The independent features are listed in table 1, 2 and 3. We built five models, each with a different set of independent features: **(1)** only time-based features; **(2)** only strictly implicit features; **(3)** only features based on content ordering; **(4)** a combination of strictly implicit features, features based on content ordering and some additional features which combine implicit feedback and content

Table 2. Content features describing the structure of the newspaper and additional features originating from combining the strictly implicit & content features.

Feature Name	Description
Content features	
pageNumber	The page number of the current page.
isFirstPage	Whether the current article is on the first page of the newspaper, or not.
isLastPage	Whether the current article is on the last page of the newspaper, or not.
category	Each article has an associated category.
catPageNumber	The sequential order of presenting the page, calculated by category.
nrWordsArticle	The number of words of the article.
nrWordsPage	The total number of words on the page.
articleIsAd	Whether the article is an advertisement, or not.
nrImgsOnPage	The number of images on the page.
nrArtsOnPage	The number of articles on the page.
pageHasTeaser	Whether the current page has a teaser to another page, or not.
articleIsTeaser	Whether the article is a short teaser which links to another article, or not.
isTeasedArticle	Whether the current article was teased earlier in the newspaper, or not.
isTeasedPage	Whether the current page was teased earlier in the newspaper, or not.
nextPageNrWords	The number of words on the next page.
prevPageNrWords	The number of words on the previous page.
Implicit & content features combined	
swipeFreqArtWords	Swipe frequency by every 100 words of the article (100×nrSwipesArticle/nrWordsArticle).
swipeFreqPageWords	Swipe frequency on the page by every 100 words on the page (100×nrSwipesPage/nrWordsPage).
swipeDevArticle	The deviation in number of swipes from the average number of swipes on an article for this user.
swipeDevPage	The deviation in number of swipes on the page from the average number of swipes on a page for this user.
swipeDevPageNr	The deviation in number of swipes on the page from the average number of swipes on a page with this page number for this user.

ordering information (see table 2); **(5)** all features combined (this includes again some additional features which combine implicit feedback and dwell time information, see table 3). These features were selected based on two major considerations: which features could be created based on the raw data the modified apps collected, and by synthesizing previous research on the usefulness of device interactions, as discussed in the literature review. Some features relate to 'teasers'. A teaser is a small image or piece of text on a page of the newspaper which links to another article in the same newspaper. It is often positioned by the editors on the first pages of the newspaper, and is there to persuade users to read other articles further in the newspaper.

We evaluate the predictive power of the models by calculating the AUC on the test set. We show ROC curves which plot the true positive rate against the false positive rate, and use the

DeLong et al. [1988] test to assess whether the AUC of two models is statistically different. We report the sensitivity and specificity for that threshold which yields the largest value for the Kolmogorov-Smirnov statistic between the predictive model and a random prediction model. The Kolmogorov-Smirnov test is a non-parametric test which measures the maximum difference between two cumulative distribution functions [Lilliefors 1967].

Showing the ROC curves, the AUC scores, and the sensitivity and specificity allows us to compare the predictive performance of the five different groups of independent features, thereby answering the first part of the first research question.

To answer the second part of the first research question, we report the top five most important features of each model by ranking each of the features on the p-value of the Wald statistic, the logistic pseudo partial correlation, the adequacy and the c-statistic (calculated over the whole dataset), and then taking the average of these four rankings to produce a final importance ranking for each feature, as in [Harrell 2015].

Besides showing the five highest ranking features for each model, we also calculate the odds ratio *ceteris paribus* of the top five features of each model. In logistic regression, the odds ratio of an independent feature describes the multiplicative increase in the odds of the dependent feature given a one-unit increase in that independent feature. It is calculated by exponentiating the logistic model coefficients. In this study, the odds ratio of a feature in one of the models describes the change in the odds of a user being engaged with an article, given a one-unit increase in the feature value. Odds ratios can be used to compare the magnitude of the effect of different independent features. An odds ratio larger than one is associated with higher odds of a user being engaged, while an odds ratio smaller than one is associated with lower odds of engagement occurring [Baesens et al. 2015]. Using odds ratios makes interpretation of effect sizes easy and can be used to communicate with non-technical content creators. However, sometimes we have to be careful when interpreting these odds ratios, because we observe some correlation between the independent features, which makes interpreting the odds ratios *ceteris paribus* harder. When we present and discuss the most important features in the results, we always mention when a feature is highly correlated with another feature.

The second research question is also answered by showing the predictive performance of the models which include features based on content ordering as independent features and contrasting their performance with the other models.

To answer the third research question, about briefly read content, we restrict the observations to only keep those user-article pairs on which at most 15 seconds were spent. This threshold was chosen together with the newspaper editors. For this subset of observations, we also report the AUC, the ROC curves and the odds ratios of the most important features. This allows us to contrast the usefulness of the different groups of features when we limit the observations to only briefly read content.

For the final research question, concerning the most frequently read articles, we subset the data on the top 25% of articles which were read by the highest number of users. Again, we also report the AUC, the ROC curves and the odds ratios of the most important features.

4 EXPERIMENTAL SET-UP

We did experiments with people who use a tablet app for reading a digital newspaper. We included two separate newspaper brands which are published by the same mother company. Respectively 198 and 209 paying subscribers of each newspaper who used the app regularly participated in the experiment. Each of the two experiments had a duration of one month. The newspapers' brand names can not be mentioned due to confidentiality reasons, but they exist already for several years

Table 3. Additional features originating from combining of time-based and strictly implicit features.

Feature Name	Description
Implicit & time-based features combined	
swipeFreqArticleTime	Swipe frequency by each minute spent on the article (60×nrSwipesArticle/timeOnArticle).
swipeFreqPageTime	Swipe frequency by each minute spent on the page (60×nrSwipesPage/timeOnPage).
pageNrReadProb	The probability (calculated over all users) saying whether the page with this page number will be read or not.
persPageNrReadProb	The probability for this user with which the page with this page number will be read or not.
catReadProb	The probability (calculated over all users) saying whether the category to which the current article belongs, will be read or not.
persCatReadProb	The probability for this user saying whether the category to which the current article belongs, will be read or not.
catPageNrReadProb	The probability (calculated over all users) with which the category to which the current article belongs, will be read or not, taking into account the sequential order of presenting the page, within a category.
persCPgNrReadProb	The probability for this user saying whether the category to which the current article belongs, will be read or not, taking into account the sequential order in which the page is presented, within a category.
devMeanTimeOnPage	The deviation of the average time on a page for this user.
devMeanTOPageNr	The deviation of the average time on a page for this user, taking into account the current page number.

and have each more than 10,000 active users. Both newspaper brands are among the five most popular in a Western-European country.

We emailed a selection of paying subscribers of each of the newspapers with an invitation to fill in a recruitment survey, which assessed eligibility for participation in the experiment. The survey consisted of sociodemographic questions and questions concerning the user's typical reading behavior. Based on the answers to this survey, a sample of candidate participants was drawn which was representative for each newspaper's population of subscribers. All of our candidate participants were acquainted with the app and used it regularly (at least weekly, often more frequently). This set of candidate participants received a personal invitation to download and use the modified version of the app during the next month.

All users who participated in the experiments explicitly agreed to give in-app feedback regularly. The users were asked, before being allowed to participate in the experiment, whether they would want to indicate being engaged on a regular basis during one month by using the blue dots. Before the start of the one month experiment period, the users were also informed about the purpose of the research. We gave an explanation about the definition of engagement, clarified how the in-app feedback mechanism worked, and described which other behavioral interaction data the modified app collected. Our explanation of user engagement followed Lalmas, O'Brien, and Yom-Tov [2014], and included a definition of user engagement and overview of the associated attributes. Participants were instructed to give explicit feedback on those articles they found engaging while they were reading, but also to continue using the app and reading the newspaper like they normally would.

Fig. 1. In-app screenshot (anonymized) of a random page of one of the two digital newspapers. The blue dots are present next to every title, image and at the end of each article.

No further instructions were given to minimize participation bias. By consequence, we can assume that all users were properly motivated to give explicit feedback.

Before each of the two quantitative experiments, we performed two qualitative pilot studies, one for each newspaper brand, with about 30 readers of each newspaper brand (so 60 people in total over the two pilot studies). The pilot studies consisted of two-hour long in-depth interviews with each participant, conducted by an independent experienced market researcher who was appointed by the newspaper companies. These participants were not part of the later experiments. One of the main goals of the pilot studies was to select a good measurement methodology by letting users give feedback on different possible implementations of the in-app feedback mechanism.

We worked together with Twipe (`www.twipemobile.com`), the company which developed the apps, to modify the apps for the experiment to include an in-app feedback mechanism. An example of an anonymized screenshot of the app can be seen in figure 1. We added small blue dots next to the title, images, and at the end of the text of each article to give readers plenty of opportunities to give feedback. When a user taps one of the blue dots, a pop-up appears where a thumbs up or thumbs down can be given. We consider a user to be engaged with an article when she taps one of these blue dots and gives a thumbs up on that article.

The blue dots were placed strategically to minimize how they hamper reading the article. They were designed to offer a balanced level of distraction. On one hand, we wanted the blue dots to

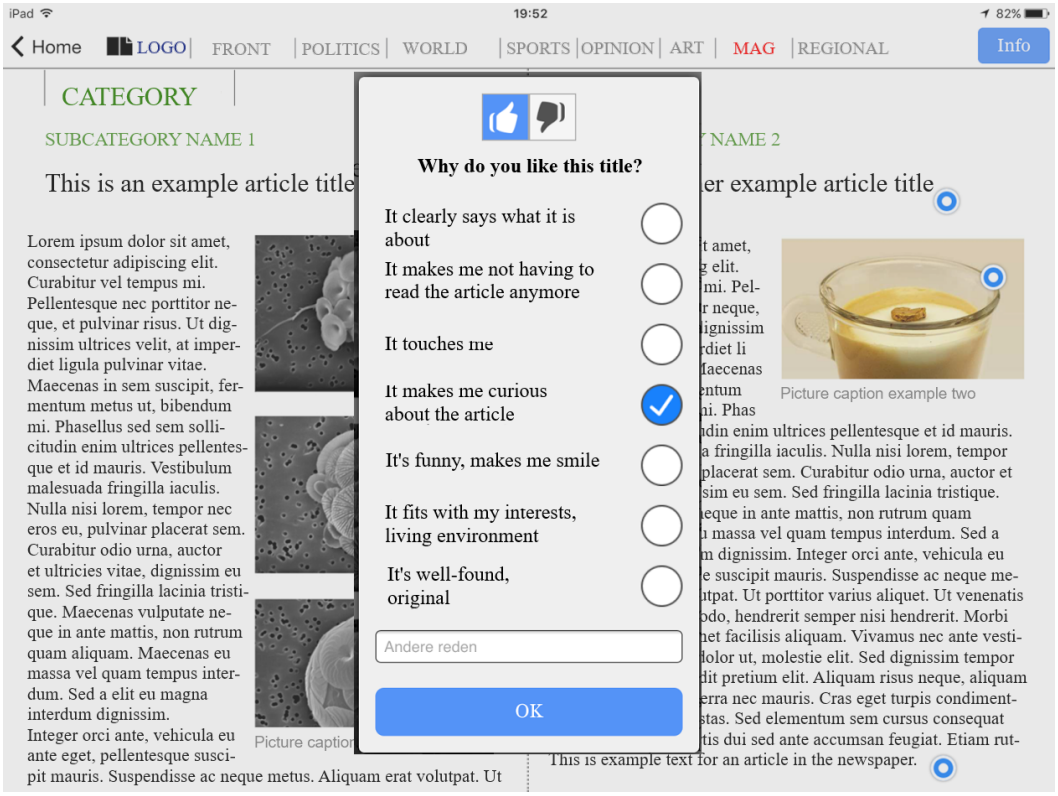

Fig. 2. Example of the pop-up that appears when the blue dot of the title is tapped. The reasons that users gave were not included in this study.

subtly attract the readers' attention, so users would remember to give feedback. On the other hand, we intended for the blue dots to be minimally distracting, so that they would not deteriorate the reading experience.

We learned during the pilot studies that it is necessary for each article to have multiple locations on the tablet interface which users could use to indicate engagement. If we would have used only one location for each article, then users would need to swipe around just to be able to give explicit feedback. For example, if we would have only a blue dot next to each title, a user who has finished reading the article, would need to swipe back up to the title of the article just to indicate that she was engaged. This would have increased the effort users would have to make. We wanted to keep the barrier to collect engagement evaluations as low as possible, as explained in the section on the methodology.

The blue dots also needed to be always readily available to the reader, they could not be hidden in a menu for example. We also decided that it would be beneficial for the reading experience to have the blue dots at the same location for every article. This way, users learn where the blue dots will be.

Taking these requirements into account and discussing this with users during the pilot studies, we decided that for each article, there was going to be a blue dot next to the title, image, and at the end of the article text. The blue dot next to the title is there to allow users to indicate feedback

quickly, even when they have read only the title or just skimmed the first paragraph of the article. Users could use the blue dot next to the images because article images often attract attention to the article even before the title is read. Also, sometimes images are shown further down the article and not immediately visible, so having a blue dot next to each image makes sure users do not have to scroll back up to the title of the article to give feedback. The blue dot at the end of the article text makes sure that when users finish reading the article, they do not have to scroll back up to give explicit feedback, which would also contaminate the behavioral interaction variables.

Before the experiments started, we explained to users that all feedback on an article counts equally, irrespective of where they tapped the blue dot. We informed the participants that during the analysis afterwards, their feedback would be summarized as engagement with the article as a whole because that is the most relevant to the newspaper creators. More concretely, if one user gives a thumbs up to an article by tapping the blue dot next to an image of that article, and another user gives a thumbs up to the same article by tapping the blue dot next to the title of the same article, we count both users as both having been engaged with that same article. Sometimes users give feedback on more than one aspect of the article (e.g. give a thumbs up to both the title and the image of the same article), but we take these interactions together as one observation and consider a user to be engaged when she gives positive feedback about at least one aspect of the article. We did not instruct the participants to use either the blue dots next to the title, image or at the end of the article depending on which element of the article they were most engaged with. However, we acknowledge that despite these instructions, users could still decide to use a blue dot in a different location to indicate a different type of engagement.

Normally, users would not want to avoid the blue dots because they explicitly opted in to participate in the experiment. Nevertheless, we acknowledge the possibility that during the quite long experiment period of one month, users might have started to forget or have gotten tired of giving explicit feedback during each reading session. We did mitigate this situation by monitoring the number of times each user was giving feedback on a daily basis relative to their total time spent in the app, and sending short reminder emails when we noticed a significant drop in the number of times feedback was given. We sent at most three reminder emails to the same user, spread out throughout the one month experiment period.

A user indicates being engaged by tapping one of the blue dots and giving a thumbs up. When a user taps one of the blue dots, only the thumb up or down is visible. Only after a user taps the thumbs up, a number of reasons for giving this explicit feedback become visible in a pop-up, as can be seen in figure 2. Selecting an reason for being engaged in the pop-up was optional, users could also just continue reading after giving explicit feedback and frequently did so.

The items in the pop-up are derived from summarizing the feedback from the participants in the qualitative pilot studies. Five out of the seven items are the same for the two newspaper brands. The items in the pop-up are very short phrases and there are only seven, so it does not take long to select one, if the user wanted to. As selecting one of the reasons in the pop-up is completed in a matter of seconds, it is not a large cognitive load and thus requires very little user effort. The items are also the same for every article, so users could learn very rapidly what the possible items were, as they gave feedback many times during each reading session, and the experiment lasted for a month.

Our decision to include a pop-up allows users to optionally indicate the reasons for why they were engaged. The multiple reasons which could be selected are not included in the analysis done for this paper. In the current paper, we use a simple, binary evaluation of engagement, in order to stay as close as possible to how the normal reading experience would occur in a natural setting, without assigning much further meaning to it or hypothesizing about why the user was engaged.

Feedback from participants in the pilot studies confirms that this light-weight mechanism allows us to add some meaning to engagement, in such a way that is not too disruptive to the regular reading experience users are used to.

However, the incorporation of these reasons in the pop-up in the in-app feedback mechanism is a limitation for the current paper. The regular reading experience would be more like it would occur in an uncontrolled environment if the optional reasons in the pop-up would not be present.

The decision to include these items in the pop-up in the in-app feedback mechanism was motivated by the fact that the current paper is built on experiments which are part of a larger research project, executed in collaboration with publishing companies, which investigates the trade-off between accuracy and scalability for different ways of measuring user engagement. One of the goals of the broader research project is to better understand engagement in the specific context of news reading.

Including the reasons in the pop-up in the in-app feedback mechanism permits us to do future research where we analyze the items users selected when they indicated that they were engaged. We could for example inspect whether interaction behavior on the tablets is different depending on the different reasons users give, or analyze whether people give different reasons for being engaged depending on whether they used the blue dot next to the title, image or at the end of the article to indicate their engagement.

To get more accurate measurements for the time a user spent on an article, the time spent between tapping a blue dot to give feedback and tapping *OK* which signified the end of giving feedback, was subtracted from the total time spent on the article.

Users can also give a thumbs-down on an article, but this occurred very infrequently and primarily happened on advertisements in the newspapers. By consequence, we excluded all observations where the article was rated negatively by the users.

A particular aspect of the digital newspaper reading experience is that the content is presented and consumed in a linear way. Users start at the first page of the newspaper and most of them swipe through the pages sequentially until they end their reading session (e.g., they swipe through page 1, 2, 3, ..., 10 and then stop reading). This linear reading aspect of newspapers implies that content ordering features might be useful for predicting engagement, and we investigate this effect in the results.

There were also a number of observations for which the calculated time spent on the article was very low or almost zero. As can be seen in figure 1, the situation could occur where multiple articles are visible in the viewport of a user at the same time. The user could be interacting with the article on the left, swiping up and down, and then all of a sudden swipe once on the article on the right. If this happens, the calculated total time spent on the article on the right is very low. This situation also occurs frequently in practice with other apps and websites: there are often links to other content, with corresponding images and multiple sentence captions next to or under the current article. This is an aspect of the experience which we could not control or mitigate.

The time spent on a page (the feature timeOnPage) is calculated as the total time between the moment the user arrives at that page until the user leaves that page. The time spent on an article (the feature timeOnArticle) is only counted from the moment a user starts interacting with the article, until the user either starts interacting with another article on the same page, or leaves the current page. If there were no interactions with a page, and there was only one article visible on that page, we assigned the time spent on that page to the time spent on the article on that page. If there were no interactions with a page, and there was more than one article visible on that page, we did not attribute any time to a specific article because we could not be sure which article the user was reading.

Table 4. Contingency table of all observations used for the main models, describing whether the article is considered to be engaging or not. Total number of observations is 59875 for newspaper A, and 48659 for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	42673	17202
% of total	71.27%	28.73%

Newspaper B

	not engaging	engaging
nr. observations	28475	20184
% of total	58.52%	41.48%

The alternative option was to divide the total time spent on the page by the number of articles which were visible on that page and to add that time to the time spent on each article on that page. However, we did not do this because we have no reason to think that the assumption that users divide their time equally over all visible articles is true or more reasonable than other explanations of how users divide their attention over all visible articles on a page.

Eventually, we collected useful data for 407 experiment participants in total, and ended up with over 100,000 unique article visits by users (see table 4).

5 RESULTS AND DISCUSSION

We now show the results for our models. We first show the results for the most general main models. Next, we discuss the results of the models for the briefly read content, and finally we analyze the models for the most frequently read articles.

In each of the next three sections, we report each time for both newspapers a contingency table of the observations used for constructing the models, the ROC curves, the AUC scores, specificity, and sensitivity of each model, and a table with for each model the top five most important features and their associated odds ratios.

In general, the specificity is the proportion of true negatives which are correctly identified by the model. In this study, the specificity is the proportion of articles on which a user spent time but did not find engaging and which were correctly predicted by the model. The sensitivity is the proportion of observations for which the model correctly predicts that there would be engagement, i.e. that a user would indicate her appreciation of the article. We evaluate the models on their AUC as it is a good summary measure of predictive performance [Baesens et al. 2015].

When discussing the most important features and their associated odds ratios (OR), we only call attention to the particularly interesting or unexpected results. Generally, the differences in feature importance which determine the rankings are minuscule. Note that the OR always need to be interpreted *ceteris paribus*. When the features we discuss are highly correlated with other features and consequentially make the interpretation more difficult, we mention this in the text. The full correlation matrix of all the features for each of the models is available in the Appendix.

5.1 Main models

The **ROC curves** in figure 3 and figure 4 immediately show that using all features yields the best predictive performance on our test set, generating an AUC of 87.96% and 81.63%. We notice a jump in the curve for the model which uses time-based features. This happens because there are a

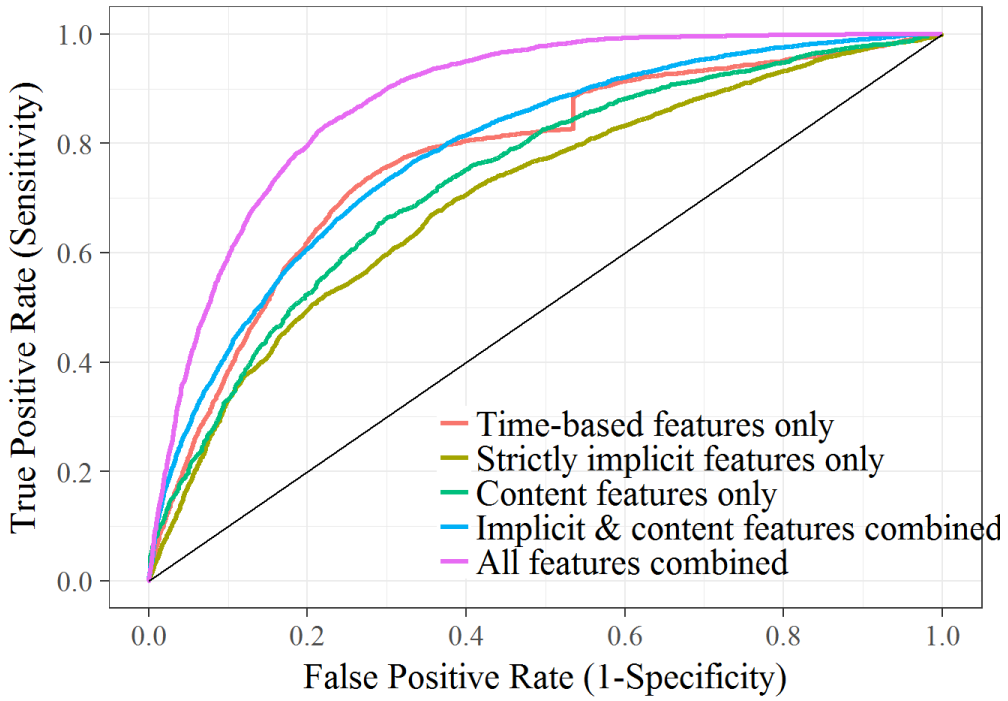

Fig. 3. ROC curves for the main models for newspaper A.

number of observations which have a very small amount of time spent on the article, as explained more thoroughly in the experimental set-up.

The **AUCs** are reported in table 5, together with the specificity and sensitivity of the predictions. At first sight, it seems like the combination of implicit & content features yields an almost equally powerful model as the model that uses only time-based features, with an AUC of 78.64% vs. 77.15% for newspaper A. With newspaper B, the difference is a bit more pronounced, with an AUC of 72.6% vs. 67.2%. However, the DeLong et al. [1988] test for comparing two ROC curves shows that the model with time-based features and the model with implicit & content features are for both newspapers significantly different (A: z = -2.7166, p-value = 0.0066; B: z = -8.6598, p-value < 0.0001). This result shows that by using *only* implicit feedback and content ordering features, we can better predict user engagement compared to using *only* time-based features.

The AUC of the model which uses only content ordering features is 73.63% for newspaper A and 67.24% for newspaper B. This model, which uses only static newspaper structure characteristics (see table 2), stands firm between the other models in its performance. It is remarkable that we can achieve this performance without even taking user interactions into account. This confirms that the editors' decisions about where to put which content and accounting for a linear reading pattern is crucial, because where the content is positioned can have a substantial impact on reader engagement occurring.

By combining the implicit & content features, the AUC increases to 78.64% for newspaper A and 72.6% for newspaper B, boosting the performance compared to using these features separately.

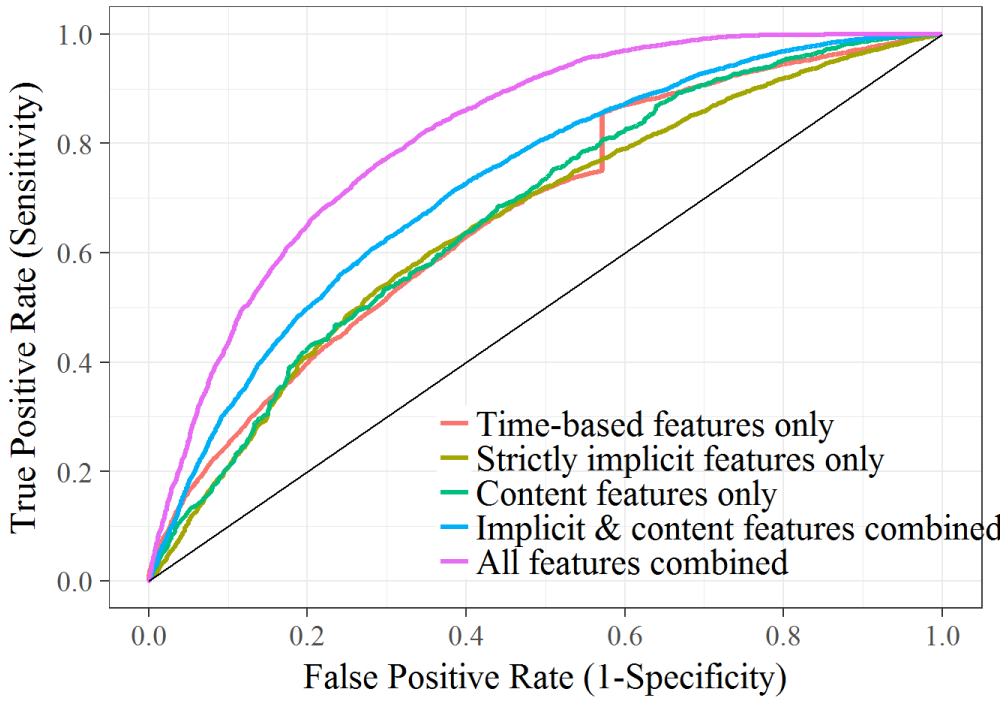

Fig. 4. ROC curves for the main models for newspaper B.

Table 5. AUC, Specificity and Sensitivity of the main models.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	77.15%	72.67%	73.51%
Implicit features only	70.4%	64.35%	66.99%
Content features only	73.63%	70.11%	66.32%
Implicit & content features combined	78.64%	70%	73.37%
All features combined	87.96%	77.41%	83.61%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	67.2%	42.85%	85.6%
Implicit features only	65.66%	68.86%	55.8%
Content features only	67.24%	56.04%	68.55%
Implicit & content features combined	72.6%	61.31%	71.69%
All features combined	81.63%	71.4%	76.1%

Table 6. This table shows for each of the main models the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

Newspaper A

	Implicit features only		Content features only		Implicit & content features combined		All features combined	
		OR		*OR*		*OR*		*OR*
1	nrSwipesArticle	1.17	nrArtsOnPage	0.85	nrArtsOnPage	0.87	swipeFreqArticleTime	0.8
2	daysSincePrevSess	1.13	category Extra	0.19	category Extra	0.17	nrArtsOnPage	0.83
			category Regional	0.2	category Regional	0.2		
			category Sports	0.35	category Sports	0.35		
3	articleCompleteness	0.99	catPageNumber	0.95	swipeDevArticle	0.8	category Extra	0.11
							category Regional	0.17
							category Sports	0.25
4	sessTimeOfDay	1.77	nrWordsArticle	1.001	catPageNumber	0.95	swipeDevArticle	0.73
5	isImageOpened	2.87	articleIsTeaser	4.65	nrSwipesArticle	1.37	nrSwipesArticle	1.5

Newspaper B

	Implicit features only		Content features only		Implicit & content features combined		All features combined	
		OR		*OR*		*OR*		*OR*
1	articleCompleteness	0.99	nrWordsArticle	1.001	articleCompleteness	0.99	swipeFreqArticleTime	0.85
2	daysSincePrevSess	1.12	nrArtsOnPage	0.91	nrWordsArticle	1.001	timeOnArticle	1.006
3	isImageOpened	1.56	category News	0.82	daysSincePrevSess	1.09	devMeanTimeOnPage	1.009
			category Econ.	0.5				
			category Sports	0.41				
			category Culture	0.35				
			category Regional	0.26				
			category Opinions	0.46				
4	nrSwipesArticle	1.15	articleIsTeaser	1.92	nrArtsOnPage	0.93	timeOnPage	0.99
5	nrSessions	0.89	isFirstPage	0.43	isImageOpened	1.42	articleCompleteness	0.996

The implicit & content features seem to complement each other, each giving information about different aspects of the user's experience.

The final model which includes all features shows that combining fine-grained swipe behavior with time spent on content gives the most additional value in terms of predictive power, as shown by the dominating ROC curve and AUC scores. It is the combination of these separate aspects of a user's experience which yields the highest predictive power for both newspapers.

We now examine the results from table 6, where we report the top five **most important features** and their corresponding odds ratios (OR).

For the model which uses only implicit features, we discuss the features which are for both newspapers related to swiping behavior (nrSwipesArticle, daysSincePrevSess, articleCompleteness, isImageOpened).

For newspaper A, for each extra swipe on an article, the odds of being engaged with that article increase by 17% (OR nrSwipesArticle: 1.17). This shows that swiping on an article is a positive indication of user interest. This confirms our intuition that more swipes in absolute numbers are strong positive indicators of engagement.

The context of the user is also important, as each day that has passed by since the user's last reading session (the feature daysSincePrevSess), the odds of judging an article positively increase by 13% for newspaper A and 12% for newspaper B. We suspect there is participant bias in play here, because we explicitly asked users to give feedback on articles during the experiment.

The feature articleCompleteness is 100% when the user scrolled down to the end of the article. Surprisingly, for both newspapers this feature has an OR of 0.99. This means that for every percentage that a user scrolls further down, the odds of judging that content positively decrease by 1%. A possible explanation is that this feature captures scanning behavior.

The most surprising feature in this model is isImageOpened. When the user taps on any image of the article to open the image and see it more clearly, the odds of being engaged with that article increase by 187% for newspaper A or 56% for newspaper B (OR isImageOpened: 2.87 and 1.56). We can conclude from this that opening an image is a behavioral action that shows clear user interest in that image, and is associated with considering the article to be engaging.

For the model which uses only content features (second column of table 6), the category feature relates to the importance of the ordering of the articles in the app. Note that for the feature *category*, the odds ratio is given for each possible level of *category* versus the base level *Front Page*. The categories are shown in the tables in the order that they appear in the app.

The feature catPageNumber for newspaper A also points in the direction of the effect that when swiping further through the newspaper, users will be less likely to judge content as being engaging.

For newspaper B, the OR of isFirstPage is low (0.43) because in the design of this app, the first page is a front cover which does not have any content which can be judged.

We observe a content ordering effect in the feature rankings, where articles in the beginning of the newspaper are on average considered to be more engaging than articles further in the newspaper. We can only hypothesize that a possible explanation for this effect is that users have a position bias where they consider the first pages of the newspaper to be more engaging. If we would want to determine whether this position bias actually exists, we would have to set up an experiment where we manipulate the position of the same articles and analyze how the engagement judgments change.

For each extra article on a page (nrArtsOnPage), the odds of being engaged with one of those articles decreases by 15% for newspaper A or 9% for newspaper B. We believe that when there are more articles visible, the user's attention is more spread out over all these different articles.

Another feature in the top five of most important features which is not related to content ordering is nrWordsArticle. For every extra 100 words in an article, the odds of being engaged increase by 10% (OR nrWordsArticle: 1.001, for both newspapers). It is a stretch to generalize this to saying that longer articles will always be more engaging for users. In our study, we find that longer articles are more likely to be considered engaging. However, this is not a causal link. This finding does for example not show that journalists should focus on writing longer articles. The relationship between the occurrence of engagement and article length probably also depends on other factors. However, we can state that very short articles have lower odds of being considered engaging.

When considering the combination of both implicit & content features, we see that the top five of these models (third column of table 6) are also present in the top five of the models with both sets of features considered separately, for both newspapers. The effect of content ordering persists with newspaper A, represented by the features category and catPageNumber.

The exception is feature swipeDevArticle for newspaper A, which is a new feature introduced by combining implicit & content features (see table 2). The OR of swipeDevArticle is 0.8, so the odds of liking an article surprisingly decrease when an article is swiped more than average. Luckily, this effect is compensated by nrSwipesArticle: just like in the model with only implicit features, for each extra swipe on the article the odds of liking that article of newspaper A increase by 37%. However, the correlation between the features swipeDevArticle and nrSwipesArt is 73%, so we can not give additional meaning to these features here. The top five features for newspaper B deliver

no new insights, they were all already encountered in the previous models, with OR's pointing in the same direction and of similar magnitude.

Finally, in the last model where all features are included, the most important feature is swipeFreqArticleTime and its OR is 0.8 for newspaper A and 0.85 for newspaper B (last column of table 6). This confirms the findings of Guo et al. [2013b] as this feature combines swipe behavior with dwell time information.

This means that when the number of swipes for each minute spent on an article increase by one, the odds of being engaged with that article decrease by 20% or 15%. When swipeFreqArticleTime is larger, there are either more swipes for the same time spent on the content or the same number of swipes for a shorter time spent. In both cases, for larger values of the swipe frequency by each minute spent on the article, the time between swipes decreases, which makes it more likely that the user was scanning the article. When a user scans an article, she is less likely to be engaged by that content. Conversely, if the values for swipeFreqArticleTime are smaller, the time between swipes increases, which means that the user was more actively reading the article. Active reading thus makes a user more likely to be engaged with the content. By combining swipe behavior with dwell time in this feature, we can infer engagement more accurately.

Both the models for newspaper A and B achieve desirable performance for predicting when a user is engaged with the content she is reading. When we consider the different groups of independent features separately, we achieve comparable predictive performance, even if we only use structural newspaper content features which do not depend on user interactions at all. The performance increases when we combine the different types of features. The results show that it is not so that there exists one group of features which is always performing better than another. It is rather the combination of different types of features that makes these models perform so well.

5.2 Briefly read articles

When we subset our data to keep only briefly read articles, time-based features are not really useful anymore for determining whether the user is engaged with the content or not. The goal here is to assess the usefulness of alternative user interactions for predicting engagement, despite the fact that she spent only less than 15 seconds with it. The proportion of engaging articles changes, as can be seen in table 7. For newspaper A, 28.11% of all articles which were judged as engaging are briefly read (4836 out of 17202 observations, see table 4). For newspaper B, this was 24.04% (4853 out of 20184 observations, see table 4). Although the class imbalance becomes stronger, this had no significant impact on the predictive results. We repeated the modeling exercise by using the SMOTE resampling technique [Chawla et al. 2002] to account for class imbalance, but the results did not differ much.

The **ROC curves** are visually shown in figure 5 and figure 6. One thing that immediately stands out is the defective performance of the model which uses time-based features, especially for newspaper B. The predictions are almost as bad as a random model. Fortunately, this confirms what we expected to see - that because we restrict the range of the possible time spent on the articles, there is less information available in features based on the time spent because of a smaller variance. The jump in the curve is very pronounced and can again be explained by the fact that there are a number of articles for which very little time spent on the article was registered (as more thoroughly explained earlier in the section on the experimental set-up).

Table 8 shows the **AUCs**. The sensitivity of this model is much lower compared to the other models for the briefly read articles. This means that the time-based features are not useful for distinguishing the engaging articles, as we expected. It is an interesting insight that including swipe

Table 7. Contingency table of the observations used in the models for the briefly read articles, describing whether the article is considered to be engaging or not. Total number of observations which are read less than 15 seconds is 35451 for newspaper A, and 18904 for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	30615	4836
% of total	86.36%	13.64%

Newspaper B

	not engaging	engaging
nr. observations	14051	4853
% of total	74.33%	25.67%

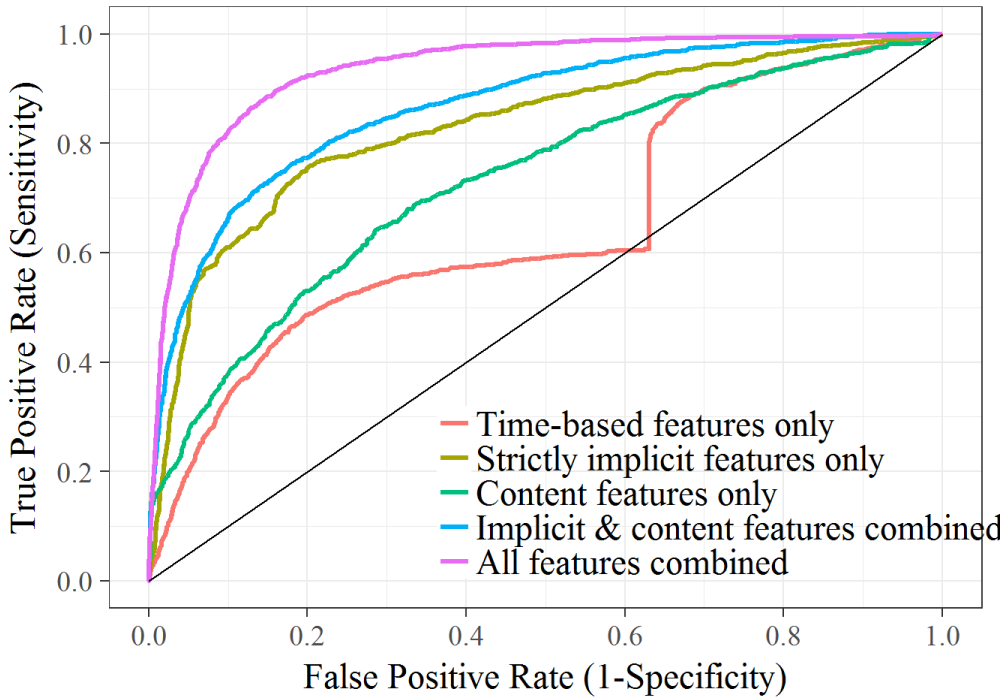

Fig. 5. ROC curves for the models for the briefly read articles for newspaper A.

interactions or content ordering characteristics is necessary for identifying engaging articles when we consider only brief interactions.

Here, the model with only implicit features performs really well with an AUC of 82.94% for newspaper A and 83.51% for newspaper B. The difference in model performance between using only implicit features and using only content features is larger compared to the models analyzed in the previous section which included all observations, and the difference in model performance between using only implicit features and using implicit & content features combined is smaller

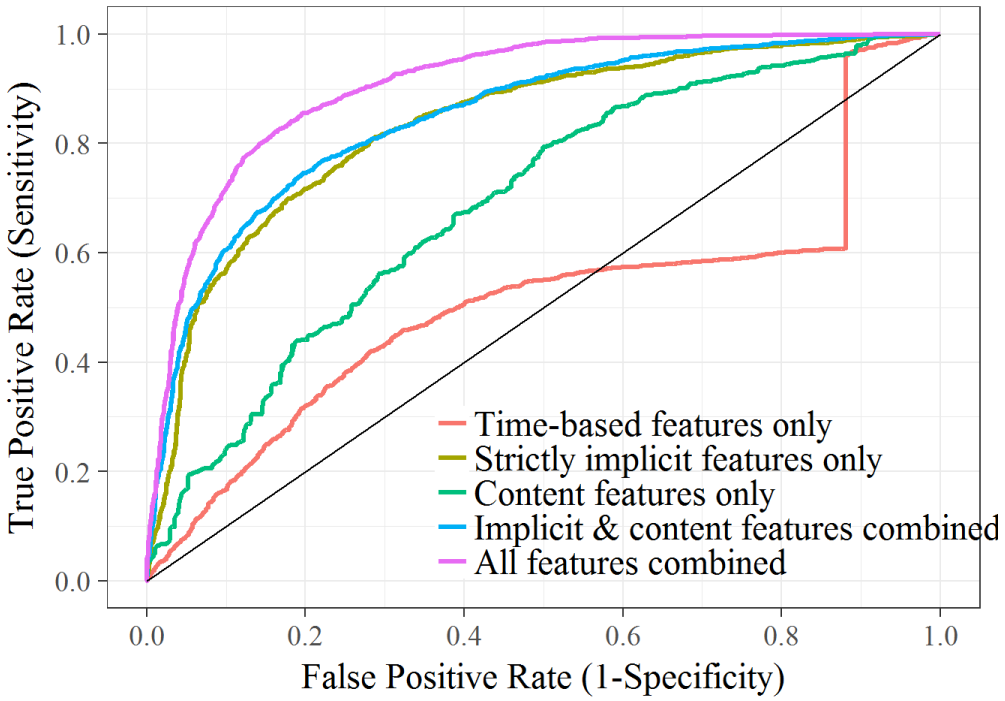

Fig. 6. ROC curves for the models for the briefly read articles for newspaper B.

Table 8. AUC, Specificity and Sensitivity of the models for the briefly read articles.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	65.1%	80.3%	48.57%
Implicit features only	82.94%	79.77%	75.91%
Content features only	73.06%	71.49%	64.18%
Implicit & content features combined	86.73%	82.24%	76.07%
All features combined	93.57%	85.19%	88.51%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	51.05%	71.83%	41.94%
Implicit features only	83.51%	71.33%	81.15%
Content features only	69.16%	50.11%	79.42%
Implicit & content features combined	84.83%	80.41%	74.51%
All features combined	90.67%	83.89%	82.1%

compared to the models from the previous section which included all observations. The ROC curves of the models which use only implicit features and the models which use both implicit & content features lie closest to each other. However, the DeLong et al. [1988] test shows that these ROC curves are significantly different from each other (A: z = 18.0007, p-value < 0.0001; B: z = 22.665, p-value < 0.0001).

The last two models which combine different types of features both perform very well. If we combine all the features described in table 1, 2, and 3, we achieve an excellent AUC of 93.57% for newspaper A and 90.67% for Newspaper B.

Based on the results of the ROC curves and the AUC scores, we can conclude that even if content was only briefly interacted with, we are able to identify engaging content by using implicit features.

Table 9 shows the top five **most important features** for each model for the briefly read articles. For the model which uses only implicit features, the top five features are exactly the same for newspaper A and B. Furthermore, for those main models from the previous section which also use only implicit features, the features articleCompleteness, nrSwipesArticle and daysSincePrevSess also appear as most important features with odds ratios pointing in a similar direction. For example, here too, articleCompleteness has an OR smaller than one, and sessTimeOfDay has a high OR. Note that there is a high correlation between the features nrSwipesArticle and nrSwipesPage, for both newspapers. We conclude that the most important features are similar to those of the corresponding model from the previous section. However, the performance of this model for briefly read articles which uses only implicit features is relatively higher compared to the best performing model which included all features. So exactly the same implicit features yield better predictive performance if we consider only briefly read articles. Those implicit features are able to compensate for the decrease in predictive performance due to the loss of usefulness of the time-based features. This model which uses only implicit features can already make accurate predictions from interactions that happen in only a short period of time.

The models which use only features based on content ordering, have four out of five top features in common and with similar odds ratios as the models from the previous section which did not restrict the reading times. Comparing newspaper A and B in the second column of table 9 shows that three out of their top five features are identical. The effect when an article is a teaser article linking to another article, is extreme for newspaper A (OR articleIsTeaser: 11.09). The model performance is almost the same relative to the corresponding main model from the previous section, so the content ordering effect is not different for briefly read articles compared to all articles. We do not observe an effect like in the previous paragraph when using only implicit features, where the same features became more useful when considering only briefly read articles instead of all articles.

When we look at the most important features for the model that combines both implicit & content features (third column of table 9), again all the features in the top five are also present with the models where both sets of features are only considered separately. The exception is swipeFreqPageWords, which is a new feature resulting from combining implicit & content features (see table 2). We have to be careful in interpreting the odds ratios here, as for example nrSwipesPage and nrSwipesArticle have a correlation of 41%. The feature articleIsTeaser has again a high OR for newspaper A, but swipeFreqPageWords has also an OR of 6.43 for newspaper B. This means that for each additional swipe for every 100 words on a page, the odds of finding the current content engaging increases by 543%. Of course, this should be nuanced when the number of words on a page is low. In this case, very little swipes are needed to achieve a high value for this feature. The majority of the top five features for both newspaper A and B relate to swiping behavior, indicating that for briefly read articles, implicit features are of more value for predicting engagement than content ordering features.

Table 9. This table shows for each model for the briefly read articles the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

Newspaper A

	Implicit features only		Content features only			Implicit & content features combined		All features combined		
		OR			*OR*		*OR*			*OR*
1	articleCompleteness	0.97	articleIsTeaser		11.09	articleCompleteness	0.97	swipeFreqArticleTime		0.89
2	nrSwipesPage	0.96	catPageNumber		0.95	nrSwipesPage	0.82	articleCompleteness		0.98
3	sessTimeOfDay	1.92		Extra	0.23	sessTimeOfDay	1.7	swipeFreqPageTime		0.93
			category	Regional	0.17					
				Sports	0.46					
4	daysSincePrevSess	1.3	nrArtsOnPage		0.84	articleIsTeaser	6.12	articleIsTeaser		9.42
5	nrSwipesArticle	1.31	pageHasTeaser		4.58	catPageNumber	0.95		Extra	0.1
								category	Regional	0.13
									Sports	0.22

Newspaper B

	Implicit features only		Content features only			Implicit & content features combined		All features combined		
		OR			*OR*		*OR*			*OR*
1	articleCompleteness	0.96	nrArtsOnPage		0.88	articleCompleteness	0.97	swipeFreqArticleTime		0.82
2	nrSwipesArticle	1.25		News	0.66	nrSwipesPage	0.76	articleCompleteness		0.98
				Econ.	0.32					
				Sports	0.32					
			category	Culture	0.16					
				Regional	0.09					
				Opinions	0.24					
3	daysSincePrevSess	1.27	articleIsTeaser		4.29	daysSincePrevSess	1.23	swipeFreqPageTime		0.91
4	nrSwipesPage	0.92	isFirstPage		0.22	swipeFreqPageWords	6.43	timeOnArticle		1.1
5	sessTimeOfDay	1.32	nrImgsOnPage		0.89	nrWordsPage	0.99	daysSincePrevSess		1.29

The final model combines all features and has a high AUC of 93.57% for newspaper A and 90.67% for newspaper B. The top three features for the model with all features combined are the same for newspaper A and B. If we look at the features that are most important in contributing to that predictive power (last column of table 9), we find again that combining the time aspect with the swiping behavior yields the two most important features, swipeFreqArticleTime and swipeFreqPageTime. These features describe the swipe frequency by time spent on the article and page, and tell us more about whether a user is scanning or actively reading (as explained earlier in the previous section).

Also notice that including time-based features in addition to implicit & content features still boosts the predictive performance a bit higher. This is surprising because we need to take into account that there is a lot less variation in the time-based features now. It is probably not the addition of the simple time-based features which causes the performance boost, but the inclusion of exceptional features such as swipeFreqArticleTime, which succeed in combining swipe interactions with dwell time in one feature.

Finally, the set of most important features for the models which use only implicit features and the models which use all available features does not vary a lot between the main models from the previous section and the models for briefly read articles. Fine-grained swipe interactions as implicit features are of great value for predicting engagement when users only briefly interact with some content.

5.3 Frequently read articles

It is interesting to look into the subset of top 25% most frequently read articles for several reasons. Newspaper editors spend a lot of time with the best performing content, analyzing why it works well and trying to replicate it with other stories. There might also be specific user interactions which are indicative of content which appeals to many subscribers of the newspaper. These interactions would help to identify engaging articles which function as a greatest common divisor across the whole reader base of the newspaper.

Table 10 shows the subset of observations of people spending time on and interacting with the top 25% most read articles. Although we retain only 25% of all unique articles present in the full dataset, these observations represent 71% of all observations for newspaper A, and 56.7% for newspaper B. There is again some class imbalance but adapting the modeling approach by employing a resampling scheme did not significantly improve the results.

The **ROC curves** in figure 7 and figure 8 show visually the predictive performance of the models for these most frequently read articles, and table 11 reports its performance in terms of **AUCs**. Here, the models with only time-based features outperform the models with implicit & content features combined, for both newspapers. For newspaper A, the model with time-based features achieves an AUC of 77.49% compared to 74.79% for the model which uses implicit & content features combined, and with newspaper B the difference is 68.43% against 65.95%. This means that for the most frequently read articles, specific fine-grained swipe interactions do not increase predictive power additionally to time-based features. This is in contrast to the results from the two previous sections, where the combination of implicit & content features in both sections outperformed the models which used only time-based features.

The model based on content features alone does not perform well with an AUC of 68.61% for newspaper A and 59.26% for newspaper B. However, for this model, this is to be expected. We selected the observations in this section based on how frequently the article was read, and those articles which are most frequently read share the same characteristics. The articles in this subset of the data are on general topics which many people find interesting, are typically very newsworthy, and are situated on the first pages of the newspaper. There is no content ordering effect with the most frequently read articles.

The best model, which uses all features, performs about 9% better in AUC compared to the second-best model, which uses only time-based features, for both newspapers. This shows that

Table 10. Contingency table of the observations used in the models for the most frequently read articles, describing whether the article is considered to be engaging or not. The top 25% most frequently read articles account for 42685 observations for newspaper A and 27605 observations for newspaper B.

Newspaper A

	not engaging	engaging
nr. observations	28902	13783
% of total	67.71%	32.29%

Newspaper B

	not engaging	engaging
nr. observations	13869	13736
% of total	50.24%	49.76%

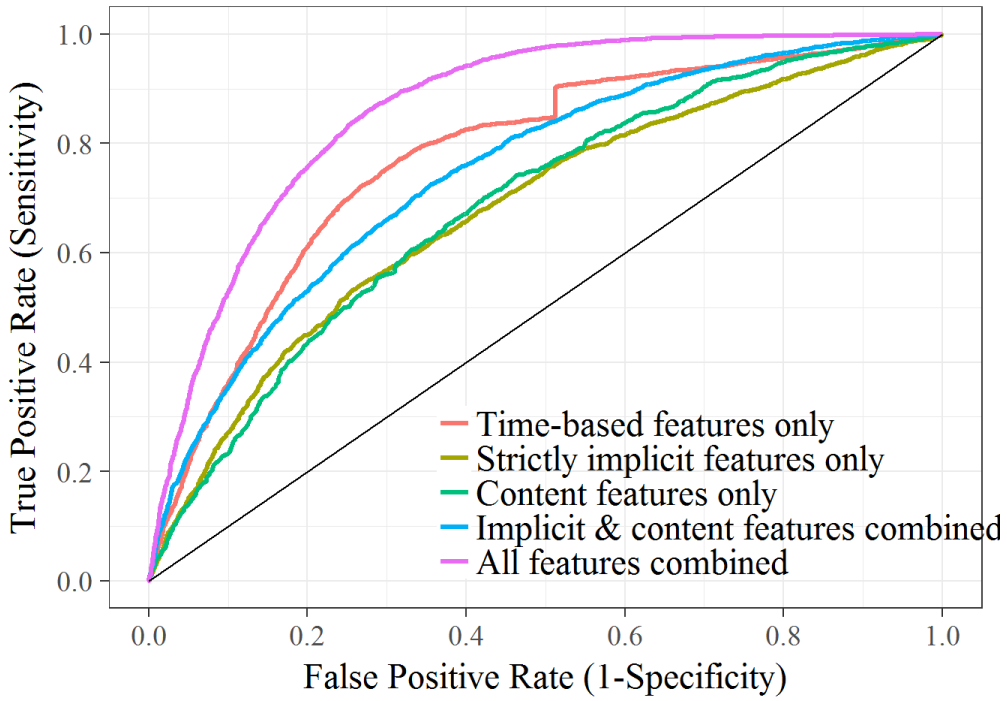

Fig. 7. ROC curves for the models for the most frequently read articles for newspaper A.

Table 11. AUC, Specificity and Sensitivity of the models for the most frequently read articles.

Newspaper A

Model Name	AUC	Specificity	Sensitivity
Time features only	77.49%	68.47%	77.11%
Implicit features only	67.91%	73.27%	54.24%
Content features only	68.61%	53.81%	74.07%
Implicit & content features combined	74.79%	64.51%	72.46%
All features combined	86.36%	72.33%	86.1%

Newspaper B

Model Name	AUC	Specificity	Sensitivity
Time features only	68.43%	54.15%	78.93%
Implicit features only	62.74%	65.84%	55.16%
Content features only	59.26%	58.07%	59.41%
Implicit & content features combined	65.95%	64.73%	59.91%
All features combined	77.33%	76.33%	73.38%

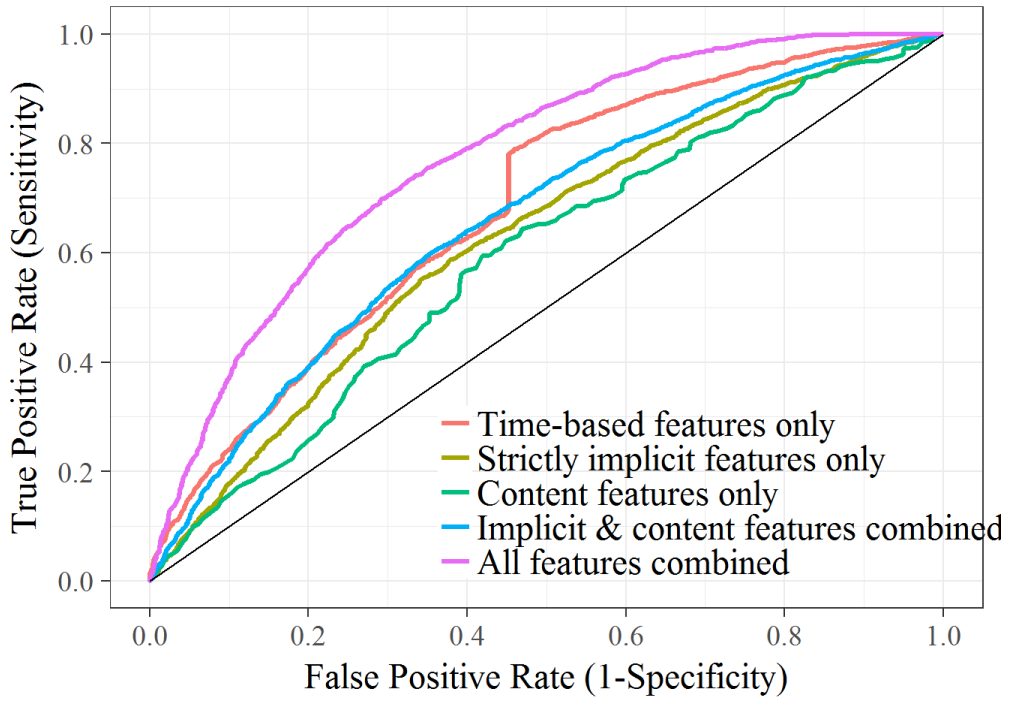

Fig. 8. ROC curves for the models for the most frequently read articles for newspaper B.

both types of features are useful in predicting engagement, and that these types of features should be used complementary. We found the same result in the two previous sections.

We now highlight some results from table 12, which summarizes the **most important features** with their odds ratios. The most important features of the model which uses only implicit features are similar to those found in the previous sections. However, the most important features of this model are not as interesting to discuss compared to the previous sections, because here, the models which use only implicit features perform weakly. The most important features of the model which uses only content-based features (second column of table 12) are almost the same as for the well-performing models with briefly read content which also used only content-based features, as discussed in the previous section. These models have weak predictive performance when we consider only the most frequently read articles. Just like subsetting on only briefly read articles in the previous section eliminated variation in the time-based features, it seems like there is now also less variation in the content-based features because we only consider the most frequently accessed articles.

Our results indicate that some, but not all features are among the most important for both newspapers. For example, features which can combine both time aspects and interaction behavior (e.g. through the swipeFreqArticleTime feature, the swipe frequency by the time spent on the article) are best for both newspapers across the three types of models we discussed. Features which appear as most important for both newspapers can offer an indication about the generalizability of the importance of these kind of features. However, we acknowledge that our study can only speculate about why some features are most important for one newspaper but not for the other,

Table 12. This table shows for each model for the most frequently read articles the top five most important features and their corresponding odds ratios (OR) ceteris paribus in the model.

Newspaper A

	Implicit features only		Content features only			Implicit & content features combined		All features combined	
		OR			*OR*		*OR*		*OR*
1	nrSwipesArticle	1.14	nrArtsOnPage		0.88	nrArtsOnPage	0.19	swipeFreqArticleTime	0.8
2	isImageOpened	3.15	catPageNumber		0.95	swipeDevArticle	0.78	nrArtsOnPage	0.83
3	daysSincePrevSess	1.10	nrWordsArticle		1.001	catPageNumber	0.95	swipeDevArticle	0.74
4	sessTimeOfDay	1.69	category	Extra	0.14	nrSwipesArticle	1.41	nrSwipesArticle	1.46
				Regional	0.43				
				Sports	0.38				
5	nrSwipesPage	0.97	isTeasedArticle		1.67	swipeDevPage	1.33	catPageNumber	0.94

Newspaper B

	Implicit features only		Content features only			Implicit & content features combined		All features combined	
		OR			*OR*		*OR*		*OR*
1	daysSincePrevSess	1.10	nrArtsOnPage		0.89	nrArtsOnPage	0.91	swipeFreqArticleTime	0.86
2	isImageOpened	1.53	nrWordsArticle		1.001	nrWordsArticle	1.001	timeOnArticle	1
3	nrSessions	0.9	category	News	0.78	daysSincePrevSess	1.07	devMeanTimeOnPage	1.007
				Econ.	0.65				
				Sports	0.37				
				Culture	0.25				
				Opinions	0.43				
4	nrSwipesArticle	1.14	nrWordsPage		0.99	isImageOpened	1.52	timeOnPage	0.99
5	articleCompleteness	0.99	isTeasedArticle		2.1	articleCompleteness	0.99	nrArtsOnPage	0.9

and why some features are among the most important for both newspapers. It might be the case that feature importance is largely determined by the specific design of the app. Also, depending on the underlying app architecture and how the app is programmed, some features might be easier to create than others. Our results show that it is important to take into account the time spent to program the features. Taking this into account and based on our results, one potential recommendation for practitioners when they are considering which features to implement could be to choose those features which combine different aspects of the user's experience into one feature, like for example the feature swipeFreqArticleTime does.

The complementarity of time-based features and implicit features also shows itself here in the most important feature of the best model which uses all features: swipeFreqArticleTime. This feature integrates both a time-aspect and a swipe interaction aspect of the user experience and comes back as a key feature in each of the three settings we discussed.

6 CONCLUSION

This paper proposed a solution to enable better large scale measurement and prediction of user engagement in the context of digital newspaper reading on tablets. We used the behavioral metric of positive in-app feedback on news articles as a proxy for engagement.

Although on a small scale users can be asked to give explicit feedback about content and that explicit feedback is an accurate measure for user engagement, it requires high cognitive effort and is not scalable. Traditionally, dwell time is used as a proxy for this explicit feedback which is usable on large scale.

We showed that by incorporating implicit feedback in the form of swiping interactions and using features based on the ordering of the content we can in general achieve better user engagement

predictions. We did an out-of-time validation of each of the predictive logistic regression models, for each model varying the set of independent features and assessing the performance on the AUC, specificity and sensitivity.

To evaluate the most important predictive features, we calculated the odds ratios after ranking the features of each model. The best features take into account the complementarity of time-based and implicit features. Features that can combine both are the most important features, such as swipeFreqArticleTime, which is the swipe frequency by each minute that a user spends on an article.

Finally, we also zoomed in on briefly read articles and the 25% most frequently read articles. We redid the analysis for the subset of observations of articles on which users spent maximally 15 seconds, and also redid the analysis by only taking into account the top 25% most frequently read articles. The briefly read articles could still be engaging for users, but time-based features were not useful anymore. Our results showed that we can predict user engagement for briefly read articles more accurately, specifically because we use the information present in the swipe interactions. If the swipe interaction information would not be available, the user engagement predictions would be worse. In contrast, for the 25% most read articles, the model which uses only time-based features performs better than models which use a combination of implicit features and content features.

In summary, we have presented the case for better large-scale predictions of user engagement by exploiting implicit feedback. In general, for the three settings we evaluated, leveraging features which succeed in combining time-based aspects and swipe interactions as implicit feedback into a single feature, always improved the performance of the predictions for user engagement.

Our study has the following limitations. We could have included more implicit features based on swipe motifs if we had also collected the coordinates of each swipe in a fine-grained manner, which Liu et al. [2015] showed to be useful for predicting search result satisfaction. We could also have opted for using degree of engagement as the behavioral measure, which previous research has shown to be more meaningful than a binary evaluation in the context of information retrieval. The reliability of our behavioral engagement measure was not tested in previous work, which has an impact on the internal validity of this study [Kelly 2009]. Ideally, we would have had the resources to include another control group in the experiments, which could have allowed us to measure the method effect of adding the feedback collection mechanism.

In future research, we could look at engagement trajectories, where we cluster users over several reading sessions and examine their engagement levels on the longer term. We could also further investigate the reasons behind the link between article length and engagement qualitatively.

REFERENCES

E. Agichtein, E. Brill, and S. Dumais. 2006a. Improving web search ranking by incorporating user behavior information. *SIGIR* (2006), 19–26. https://doi.org/10.1145/1148170.1148177

E. Agichtein, E. Brill, S. Dumais, and R. Ragno. 2006b. Learning User Interaction Models for Predicting Web Search Result Preferences. *SIGIR* (2006), 3–10. https://doi.org/10.1145/1148170.1148175

Ioannis Arapakis, M. Lalmas, B. Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M. Jose. 2014b. User engagement in online news: under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology (JASIST)* 65, 10 (10 2014), 1988–2005. https://doi.org/10.1002/asi.23096

Ioannis Arapakis, M. Lalmas, and George Valkanas. 2014a. Understanding within-content engagement through pattern analysis of mouse gestures. *CIKM* (2014), 1439–1448. https://doi.org/10.1145/2661829.2661909

Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. 2015. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection.* Wiley. 400 pages.

Nicholas J. Belkin, Michael Cole, and Jingjing Liu. 2009. A model for evaluation of interactive information retrieval. In *SIGIR '09 workshop on the future of IR evaluation*, Vol. 43. 7–8. https://doi.org/10.1145/1670564.1670567

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. https://doi.org/10.1613/jair.953

M. Claypool, P. Le, M. Wased, and D. Brown. 2001. Implicit interest indicators. *IUI* (2001), 33–40. https://doi.org/10.1145/359784.359836

Michael Cole, Jingjing Liu, Nicholas J. Belkin, R. Bierig, Jacek Gwizdka, Chang Liu, J Zhang, and Xiangmin Zhang. 2009. Usefulness as the Criterion for Evaluation of Interactive Information Retrieval Systems. In *HCIR 2009*. 1–4.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* (1988), 837–845.

A. Drutsa and P. Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. *WWW* (2015), 256–266. https://doi.org/10.1145/2736277.2741116

Georges Dupret and M. Lalmas. 2013. Absence Time and User Engagement: Evaluating Ranking Functions. *WSDM* (2013), 173–182.

Ashlee Edwards and Diane Kelly. 2016. Engagement in Information Search. In *Why Engagement Matters: Cross-Disciplinary Perspectives of User Engagement in Digital Media*, Heather L. O'Brien and Paul Cairns (Eds.). Springer International Publishing, 157–176. https://doi.org/10.1007/978-3-319-27446-1

Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems* 23, 2 (April 2005), 147–168. https://doi.org/10.1145/1059981.1059982

Q. Guo and E. Agichtein. 2008. Exploring Mouse Movements for Inferring Query Intent. *SIGIR* (2008), 707–708. https://doi.org/10.1145/1390334.1390462

Q. Guo and E. Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. *WWW* (2012), 569–578. https://doi.org/10.1145/2187836.2187914

Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. 2013a. Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance. *SIGIR* (2013), 153–162. https://doi.org/10.1145/2484028.2484100

Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. 2013b. Towards estimating web search result relevance from touch interactions on mobile devices. *CHI Extended Abstracts* (2013), 1821–1826. https://doi.org/10.1145/2468356.2468683

Frank E. Harrell. 2015. *Regression Modeling Strategies*. Springer International Publishing. https://doi.org/10.1007/978-1-4757-3462-1

Jeff Huang and Abdigani Diriye. 2012. Web User Interaction Mining from Touch-Enabled Mobile Devices. *HCIR* (2012).

Jeff Huang, Ryen W. White, and Susan Dumais. 2011. No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search. *CHI* (2011), 1225–1234. https://doi.org/10.1145/1978942.1979125

Jiepu Jiang, Daqing He, and James Allan. 2017a. Comparing In Situ and Multidimensional Relevance Judgments. In *SIGIR '17*. 405–414. https://doi.org/10.1145/3077136.3080840

Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017b. Understanding ephemeral state of relevance. In *CHIIR '17*. 137–146. https://doi.org/10.1145/3020165.3020176

Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. 1997. GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM* 40, 3 (1997), 77–87.

D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. 2014. Discovering common motifs in cursor movement data for improving web search. *WSDM* (2014), 183–192. https://doi.org/10.1145/2556195.2556265

D. Lagun and M. Lalmas. 2016. Understanding and Measuring User Engagement and Attention in Online News Reading. *WSDM* (2016), 113–122. https://doi.org/10.1145/2835776.2835833

M. Lalmas, H. L. O'Brien, and E. Yom-Tov. 2014. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 6, 4 (2014), 1–132.

H. J. Lee and Sung Joo Park. 2007. MONERS: A news recommender for the mobile web. *ESWA* 32, 1 (2007), 143–150. https://doi.org/10.1016/j.eswa.2005.11.010

Janette Lehmann, M. Lalmas, Elad Yom-tov, and Georges Dupret. 2012. Models of User Engagement. *UMAP* (2012), 164–175.

H. W. Lilliefors. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *J. Amer. Statist. Assoc.* 62, 318 (1967), 399–402.

C Liu, NJ Belkin, and MJ Cole. 2012. Personalization of search results using interaction behaviors in search sessions. In *SIGIR '12*. 205–214. https://doi.org/10.1145/2348283.2348314

Jingjing Liu and Nicholas J. Belkin. 2015. Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *JASIST* 66, 1 (may 2015), 58–81. https://doi.org/10.1002/asi.23160

Jiahui Liu, Peter Dolan, and Er Pedersen. 2010. Personalized news recommendation based on click behavior. *IUI* (2010), 31–40. https://doi.org/10.1145/1719970.1719976

Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. In *SIGIR '15*. 493–502.

Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2016. Time-Aware Click Model. *ACM Trans. Inf. Syst.* 35, 3, Article 16 (Dec. 2016), 24 pages. https://doi.org/10.1145/2988230

Shiyang Lu, Tao Mei, Jingdong Wang, Jian Zhang, Zhiyong Wang, and Shipeng Li. 2014. Browse-to-Search: Interactive Exploratory Search with Visual Entities. *ACM Trans. Inf. Syst.* 32, 4, Article 18 (Oct. 2014), 27 pages. https://doi.org/10.1145/2630420

Naresh K Malhotra, Sung S Kim, and Ashutosh Patil. 2006. Common Method in IS Research: A Comparison of Alternative Approaches and a Reanalysis of Past Research. *Management Science* 52, 12 (2006), 1865–1883. https://doi.org/10.1287/mnsc.l060.0597

Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware computing: Modelling User Engagement from Mobile Contexts. *UbiComp* (2016), 622–633. https://doi.org/10.1145/2971648.2971760

Lori McCay-Peet, M. Lalmas, and Vidhya Navalpakkam. 2012. On Saliency, Affect and Focused Attention. *CHI* (2012), 541–551. https://doi.org/10.1145/2207676.2207751

Masahiro Morita and Yoichi Shinoda. 1994. Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. *SIGIR* (1994), 272–281. http://dl.acm.org/citation.cfm?id=188490.188583

Vidhya Navalpakkam and Elizabeth Churchill. 2012. Mouse tracking: measuring and predicting users' experience of web-based content. *CHI* (2012), 2963–2972. https://doi.org/10.1145/2207676.2208705

Heather L. O'Brien. 2011. Exploring user engagement in online news interactions. In *Proceedings of the ASIST Annual Meeting*, Vol. 48. 1–15. https://doi.org/10.1002/meet.2011.14504801088

H. L. O'Brien, R. Absar, and H. Halbert. 2013. Toward a Model of Mobile User Engagement. *HCIR* (3 2013). http://circle.ubc.ca/handle/2429/45340

H. L. O'Brien and M. Lebow. 2013. Mixed-Methods Approach to Measuring User Experience in Online News Interactions. *Journal of the American Society for Information Science and Technology (JASIST)* 64, 8 (2013), 1543–1556. https://doi.org/10.1002/asi.22871

H. L. O'Brien and E. G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology (JASIST)* 59, 6 (2008), 938–955. https://doi.org/10.1002/asi.20801

H. L. O'Brien and E. G. Toms. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology (JASIST)* 61, 1 (2010), 50–69. https://doi.org/10.1002/asi.21229

Jeeyun Oh and S. Shyam Sundar. 2015. How does interactivity persuade? An experimental test of interactivity on cognitive absorption, elaboration, and attitudes. *Journal of Communication* 65, 2 (2015), 213–236. https://doi.org/10.1111/jcom.12147

Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88, 5 (2003), 879–903. https://doi.org/10.1037/0021-9010.88.5.879 arXiv:arXiv:1011.1669v3

Tefko Saracevic. 2007. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 1915–1933. https://doi.org/10.1002/asi arXiv:0803.1716

Bracha Shapira, Meirav Taieb-Maimon, and Anny Moskowitz. 2006. Study of the Usefulness of Known and New Implicit Indicators and Their Optimal Combination for Accurate Inference of Users Interests. *SAC* (2006), 1118–1119. https://doi.org/10.1145/1141277.1141542

Y. Song, Hao Ma, Hongning Wang, and Kuansan Wang. 2013a. Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. *WWW* (2013), 1201–1212. https://doi.org/10.1145/2488388.2488493

Y. Song, X. Shi, and X. Fu. 2013b. Evaluating and Predicting User Engagement Change with Degraded Search Relevance. *WWW* (2013), 1213–1223.

Maximilian Speicher and Martin Gaedke. 2013. TellMyRelevance! Predicting the Relevance of Web Search Results from Cursor Interactions. *CIKM* (2013), 1281–1290. https://doi.org/10.1145/2505515.2505703

S. Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Haiyan Jia, and Hyang-Sook Kim. 2016. Theoretical Importance of Contingency in Human-Computer Interaction: Effects of Message Interactivity on User Engagement. *Communication Research* 43, 5 (2016), 595–625. https://doi.org/10.1177/0093650214534962

S. Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Qian Xu, and Haiyan Jia. 2014. User Experience of on-screen interaction techniques: An experimental investigation of clicking, sliding, zooming, hovering, dragging, and flipping. *Human-Computer Interaction* 29, 2 (2014), 109–152. https://doi.org/10.1080/07370024.2013.789347

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response.* Cambridge University Press.

Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. Van Rijsbergen. 2005. Evaluating Implicit Feedback Models Using Searcher Simulations. *ACM Trans. Inf. Syst.* 23, 3 (July 2005), 325–361. https://doi.org/10.1145/1080343.1080347

Yingying Wu, Yiqun Liu, Ning Su, Shaoping Ma, and Wenwu Ou. 2017. Predicting Online Shopping Search Satisfaction and User Behaviors with Electrodermal Activity. In *WWW '17 Companion*. 855–856. http://dblp.uni-trier.de/db/conf/www/www2017c.html

Xing Yi, Liangjie Hong, Erheng Zhong, Nathan Nan, and Liu Suju. 2014. Beyond Clicks: Dwell Time for Personalization. *RecSys* (2014), 113–120. https://doi.org/10.1145/2505515.2505682

Received June 2017; revised December 2017; accepted February 2018